Computing for Research & the Worldwide LHC Computing Grid

Building a global large-scale
ICT infrastructure
for research data processing

Nikhef

David Groep
DACS & Nikhef
8 November 2022

Maastricht University | Department of Advanced Computing Sciences

Peter Higgs and Francois Englert at the 2013 Nobel prize press conference, Stockholm. Photo: Bengt Nyman, https://www.flickr.com/photos/97469566@N00

A ‘big science’ facility: the Large Hadron Collider at CERN

1964

VOLUME 13, NUMBER 16 PHYSICAL REVIEW LETTERS 19 OCTOBER 1964

BROKEN SYMMETRIES AND THE MASSES OF GAUGE BOSONS

Peter W. Higgs

Tait Institute of Mathematical Physics, University of Edinburgh, Edinburgh, Scotland

(Received 31 August 1964)

VOLUME 13, NUMBER 9 PHYSICAL REVIEW LETTERS 31 AUGUST 1964

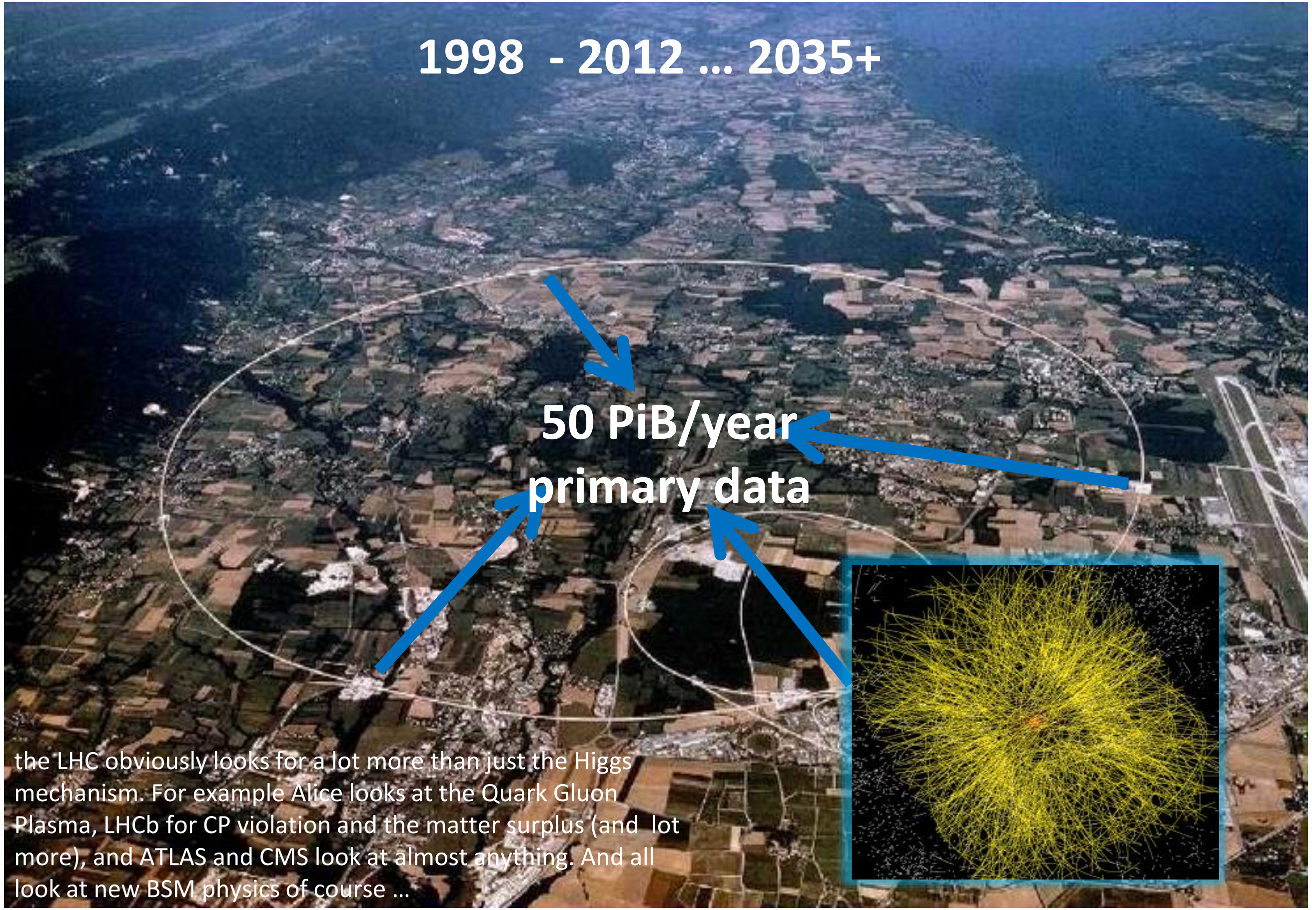
BROKEN SYMMETRY AND THE MASS OF GAUGE VECTOR MESONS*

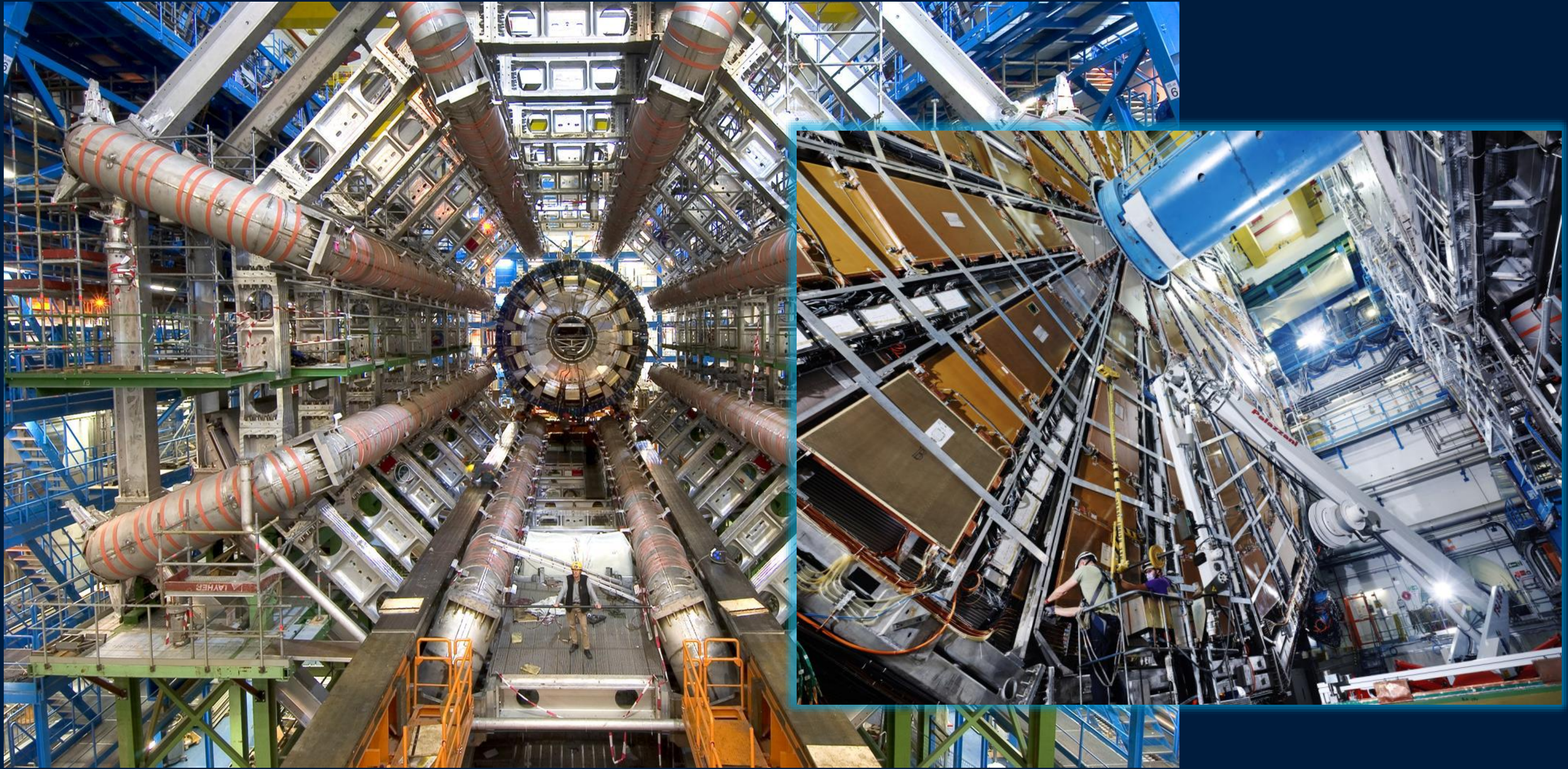
F. Englert and R. Brout

Faculté des Sciences, Université Libre de Bruxelles, Bruxelles, Belgium

(Received 26 June 1964)

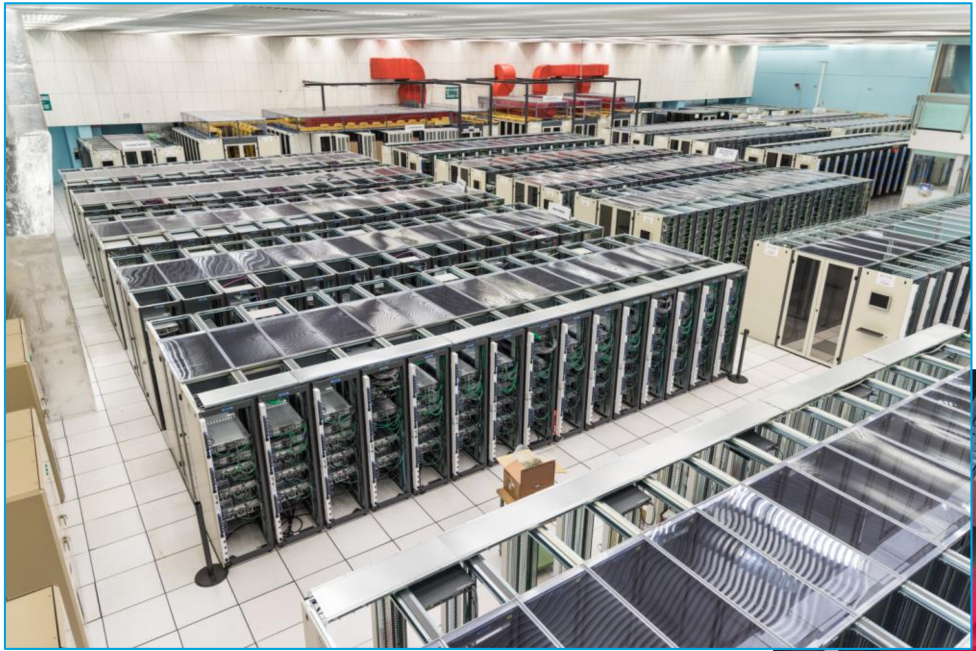
Broken Symmetries and the
Masses of Gauge Bosons
P. Higgs, Phys. Rev. Lett. **13**, 508
F. Englert, R. Brout, Phys. Rev. Lett. **13**, 321

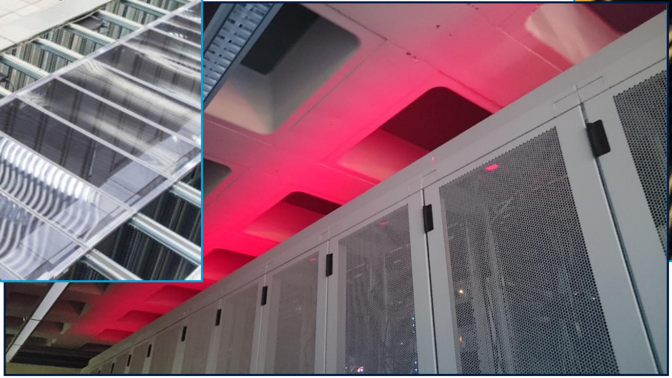


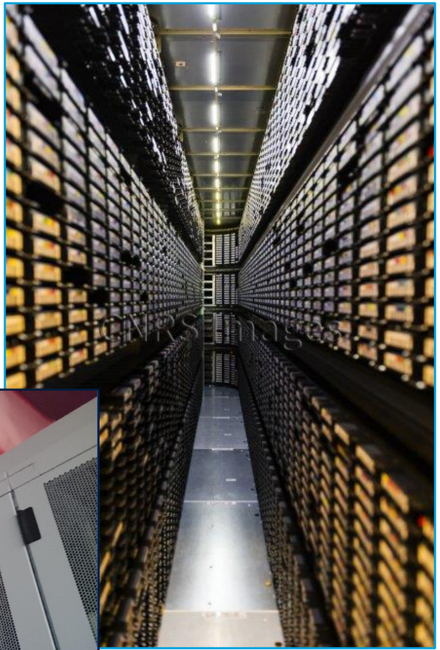


Images: ATLAS detector in the cavern at CERN. Source: CERN

‘Big Science’ needs some computing …







CERN Computing Centre B513, image: CERN, https://cds.cern.ch/record/2127440; tape library image CC-IN2P3 with LHC and LSST data; cabinets: Nikhef H234b

Our journey today … building ‘scalable’ infrastructure for the LHC computing, storage, networking and a global AAI … *if we make it to the end*

Using science use cases *from CERN’s Large Hadron Collider, the SKA radio telescope, Gravitational Wave detection, structural biochemistry (WeNMR) …*

Data intensive workflows

- the end of every faster CPUs, the thermal barrier, and the rise of parallelism

More than one …

- High Throughput Computing, herding large quantities of systems, and the cloud
- Global distributed computing, scalable storage, and data placement

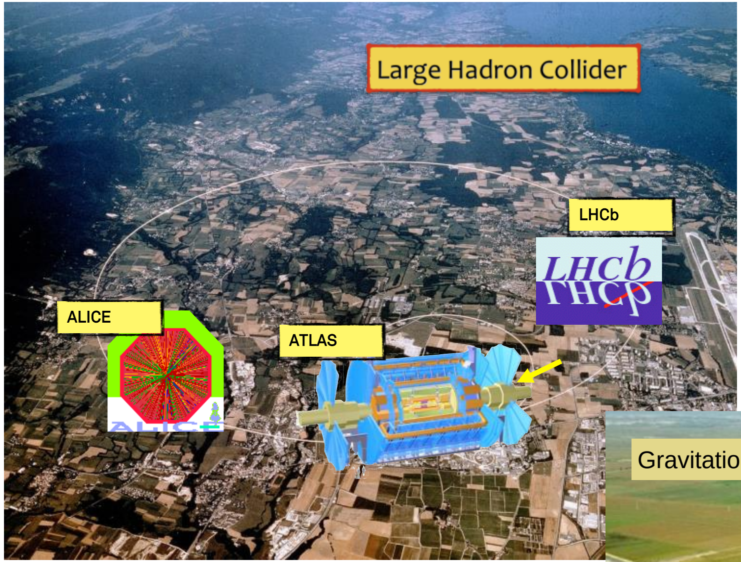
Linking ‘more than one’ into a common network

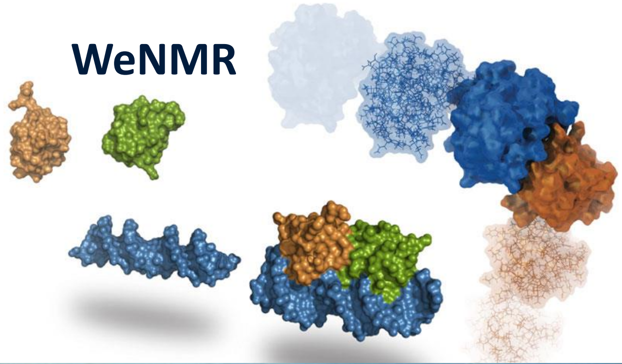
- Elephants vs. mice: shipping large quantities of data … while keeping cat videos alive
- LHC *Optical Private Network* and the *Open Networking Environment LHCone*

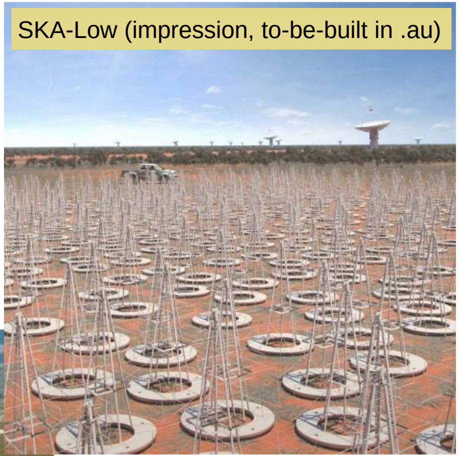
Networking the people

- Authentication and authorization technologies
- Federated identity, community management & global trust

Larger scales for both facilities and computing

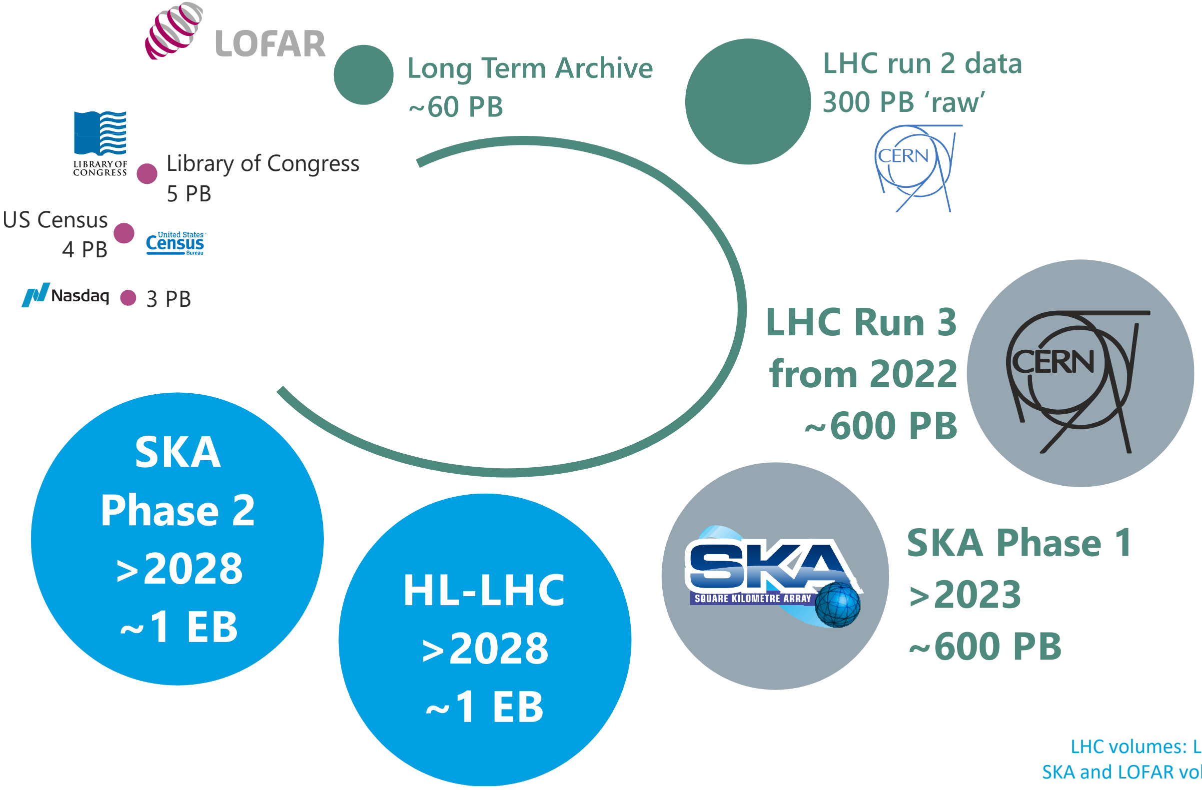






Sources: CERN https://wlcg.web.cern.ch/; HADDOCK, WeNMR, @Bonvinlab https://wenmr.science.uu.nl/; Virgo, Pisa, IT; SKAO: the SKA-Low observatory, Australia https://www.skatelescope.org/

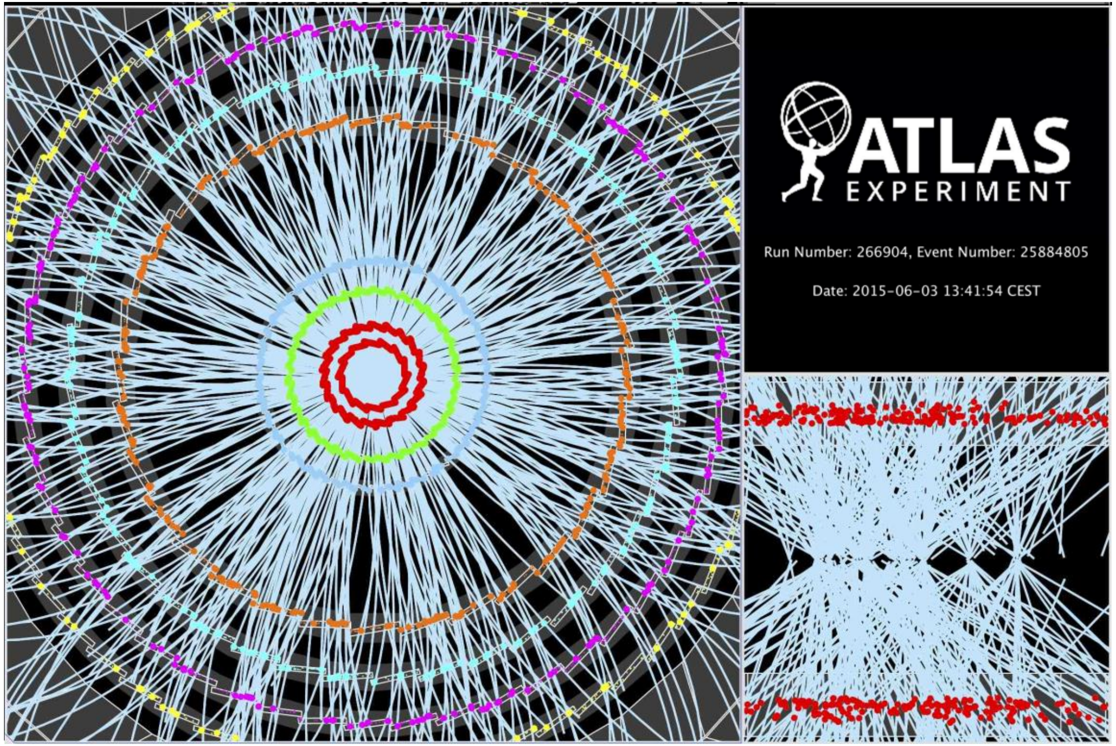
Processing at scale for data intensive science



Data from various sources, for public entities: data ca. 2018, indicative, within ~ factor 2
LHC volumes: LCG Resource Scrutiny Group & CERN; 2020
SKA and LOFAR volumes: ASTRON/Michiel van Haarlem, 2020

Computing on lots of data – 40Mevents/sec

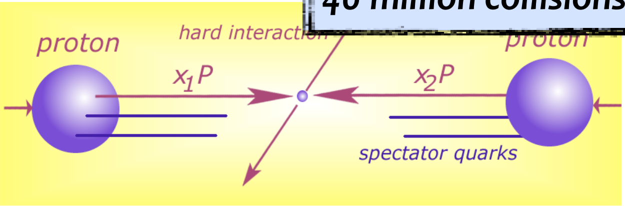
~ 10 seconds to compute a single event at ATLAS for ‘jets’ containing ~30 collisions

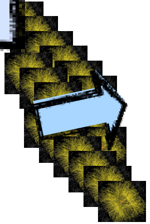


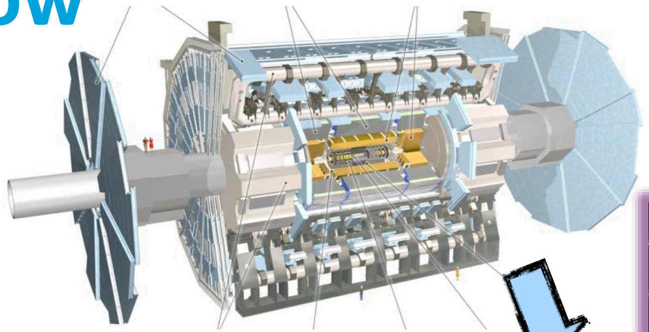
Display of a proton-proton collision event recorded by ATLAS on 3 June 2015, with the first LHC stable beams at a collision energy of 13 TeV;
Event processing time: v19.0.1.1 as per Jovan Mitrevski and 2015 J. Phys.: Conf. Ser. 664 072034 (CHEP2015)

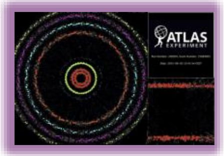
Detector to doctor workflow

40 million collisions / second

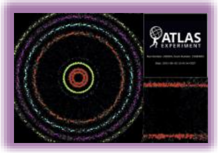








Trigger system selects 600 Hz ~ 1 GB/s data



Classify particles in collision and their physics properties:
- *electrons*
- *muons*
- *jets consisting of hadrons*

Physics analysis by (PhD) students, in papers & analysis notes

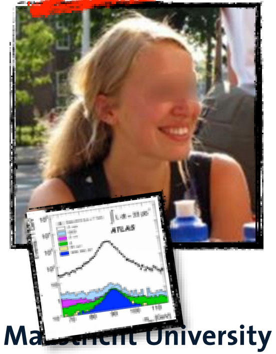


diagram adapted from Frank Linde; images: ATLAS collaboration, Nikhef. … and sorry for the GDPR-blur

Different types of large scale resources

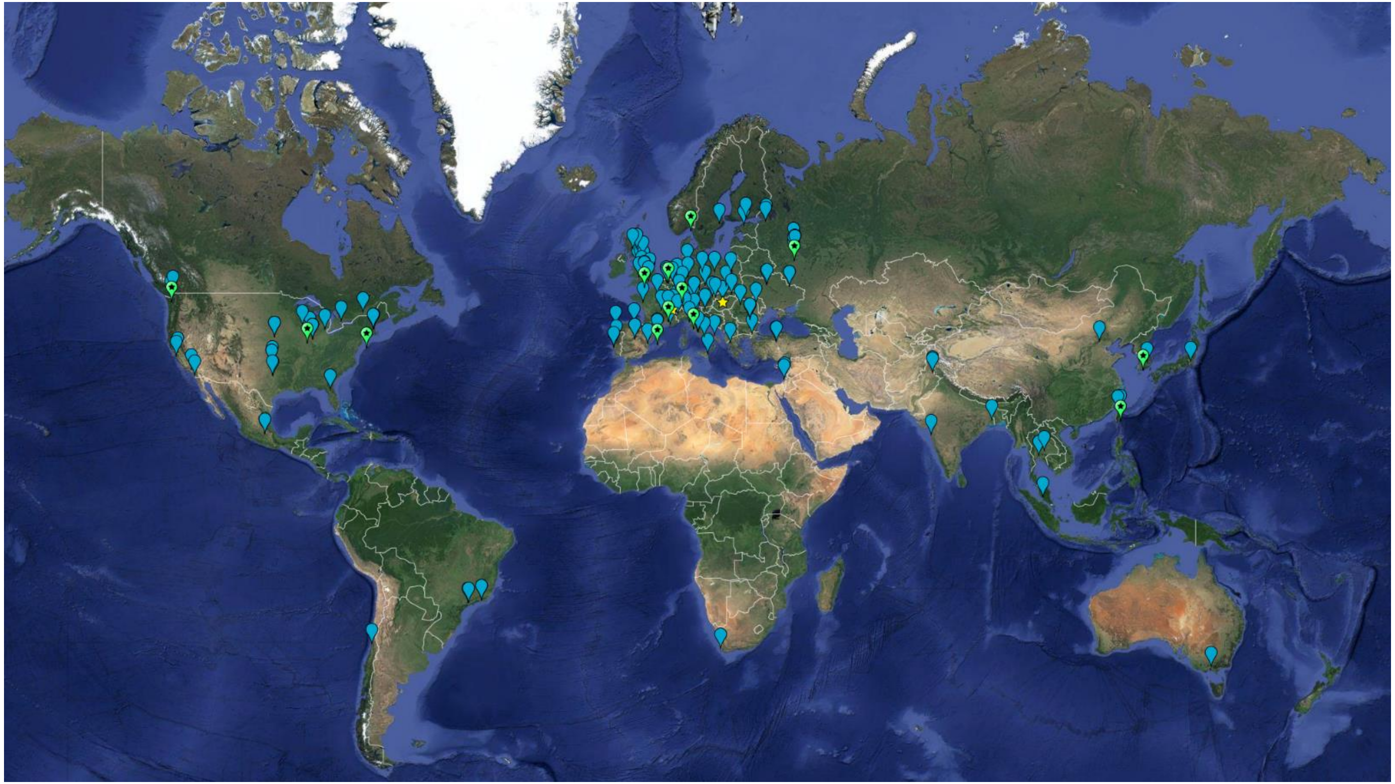
- HPC and (computational) cluster computing:
  - modelling for weather/climate, fluid dynamics, but also e.g. QC-simulation

- HTC and data-intensive processing:
  - lots of data, as in High Energy Physics (HEP), Gravitational Waves (GW) templates
  - conveniently parallel,
    but (intensive) local I/O requirements on memory and scratch storage

- portals and many web applications:
  'horizontal' scaling, possibly backed by cloud and virtualized resources
  - Cloud-native scaling and containers for 'more of the same, different each time'
  - If it's data at scale: object stores and 'CDN' web-scale caching

HPC: High Performance Computing; HTC: High Throughput Computing; K8S: Kubernetes; CDN: Content Delivery Network

Example: the worldwide LHC Computing Grid



~ 1.4 million CPU cores
~ 1500 Petabyte
disk + archival

170+ institutes
40+ countries
13 'Tier-1 sites'
NL-T1:
SURF & Nikhef

e-Infrastructures
EGI
PRACE-RI
EuroHPC
OpenScienceGrid
XSEDE (ACCESS)

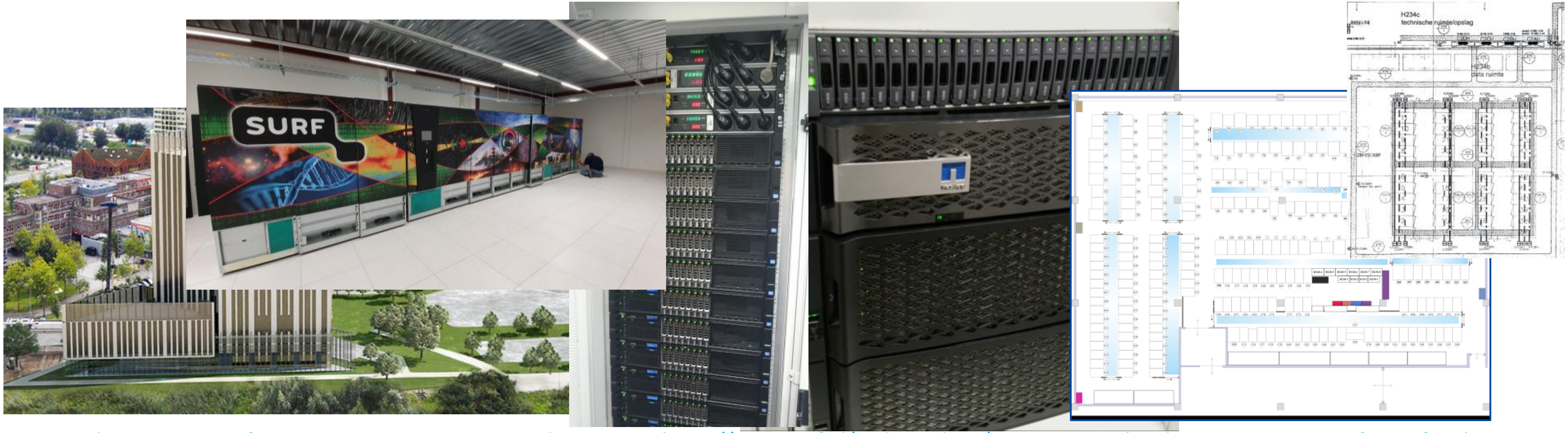
Earth background: Google Earth; Data and compute animation: STFC RAL for WLCG and EGI.eu; Data: https://home.cern/science/computing/grid
For the LHC Computing Grid: wlcg.web.cern.ch, for EGI: www.egi.eu; ACCESS (XSEDE): https://access-ci.org/, for the NL-T1 and FuSE: fuse-infra.nl, https://www.surf.nl/en/research-it

Global distribution of computing and data placement

WLCG and EGI Advanced Computing for Research

WLCG NL-T1 and the Dutch National Infrastructure

- Joint SURF & Nikhef collective service – part of EGI, WLCG and FuSE
- hosts WLCG, but also LOFAR radio telescope data, and ~100 other projects
- 59 PByte near-line storage (tape), 42.5 PByte on-line (disk), 27.6 k cores (cpu)



DNI and NL-T1 capacity from 2023 DNI NWO, LOFAR, and WLCG; see https://www.surf.nl/onderzoek-ict/toegang-tot-rekendiensten-aanvragen ; fuse-infra.nl
SURF tape total: ~80 PByte by end 2022; image library at Schiphol Rijk from Sara Ramezani; NikhefHousing: https://www.nikhef.nl/housing/datacenter/floorplan/

Single CPU scaling stopped around 2004

- limitation is power, not circuit size
  - and clock frequency is most ‘power-hungry’
  - still some packages now @ TDP of 400W
- multiple cores on the same die helped
  - AMD EPYC Genoa (Zen 4) has 96 cores on die
  - but Intel Cascade Lake AP is not even good
- CPU design-level performance gains left
  - predictive execution
  - out-of-order execution
  - on-die parallelism (multi-core)
  - pre-fetching and multi-tier caching
  - execution unit sharing (‘SMT’)

but at increased risk for security/integrity

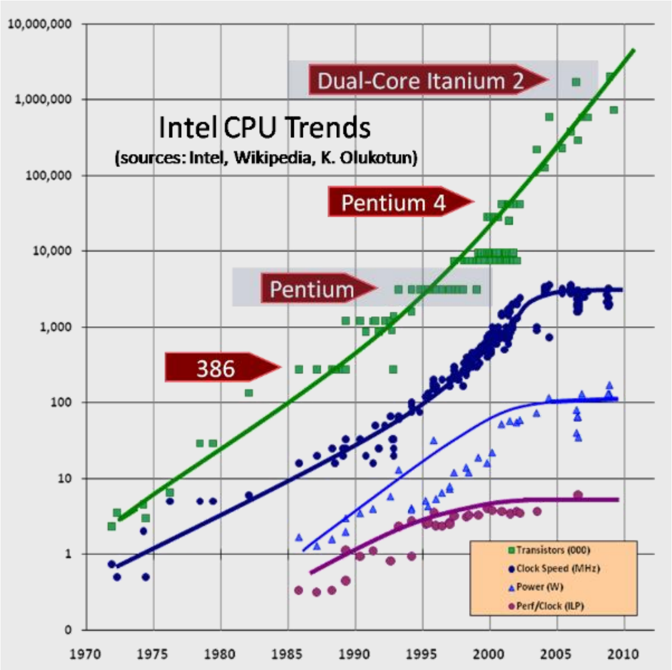
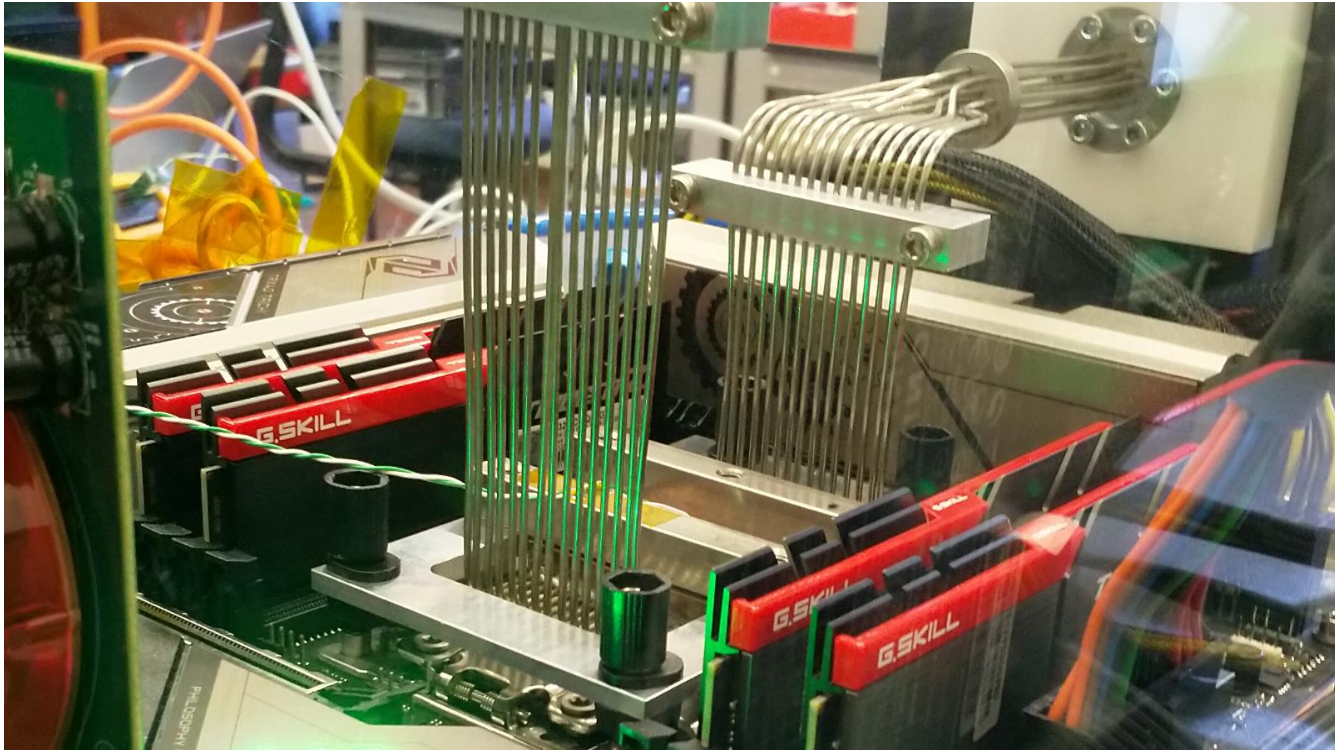


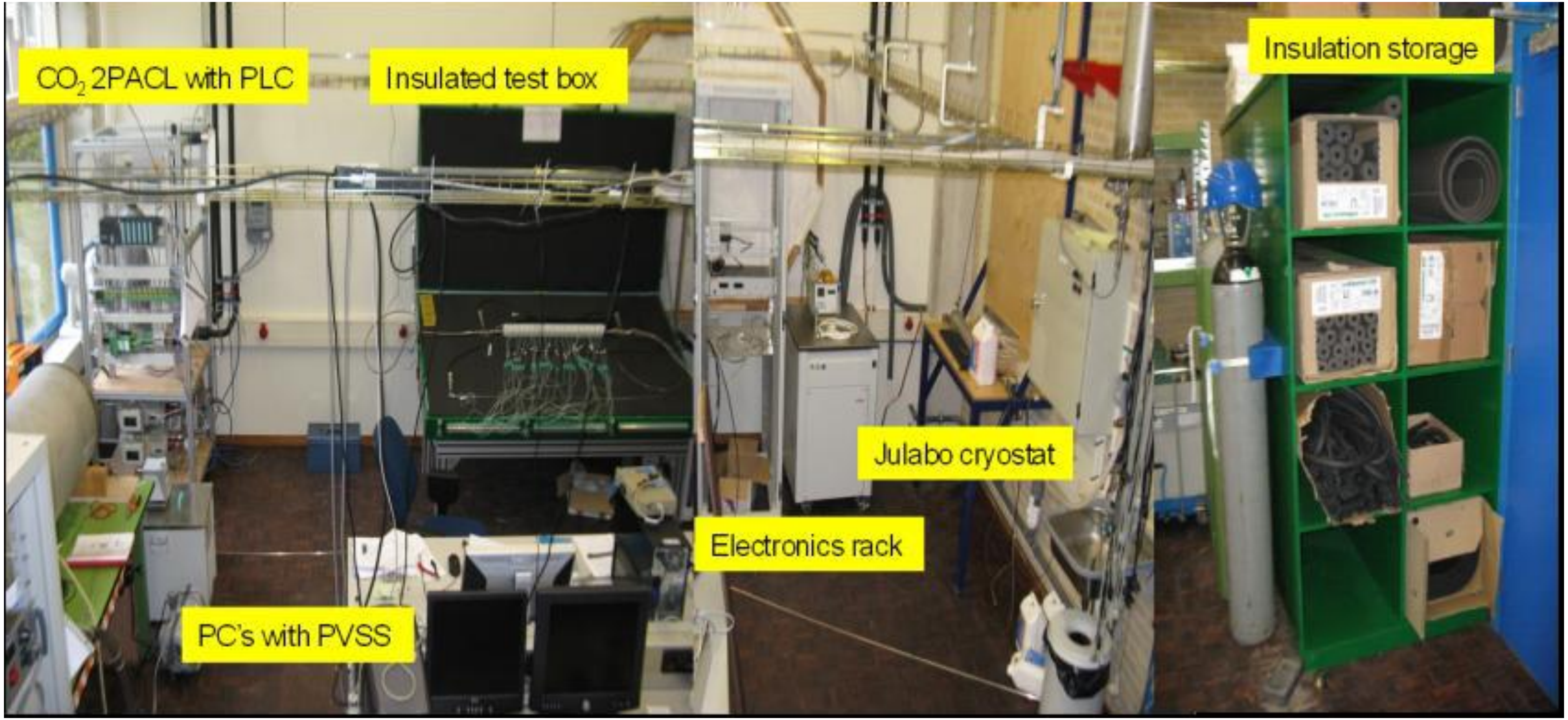
Image: Herb Sutter, *Dr.Dobbs Journal* 2004, updated 2009, see http://www.gotw.ca/publications/concurrency-ddj.htm

Fix the thing that didn't scale well, CPU frequency??



LCO2 cooling of an AMD Ryzen Threadripper 3970X [56.38 °C] at 4600.1MHz processor (~1.5x nominal speed) sustained, using the Nikhef LCO2 test bench system (https://hwbot.org/submission/4539341) - (Krista de Roo en Tristan Suerink)

… since you then need this around it …

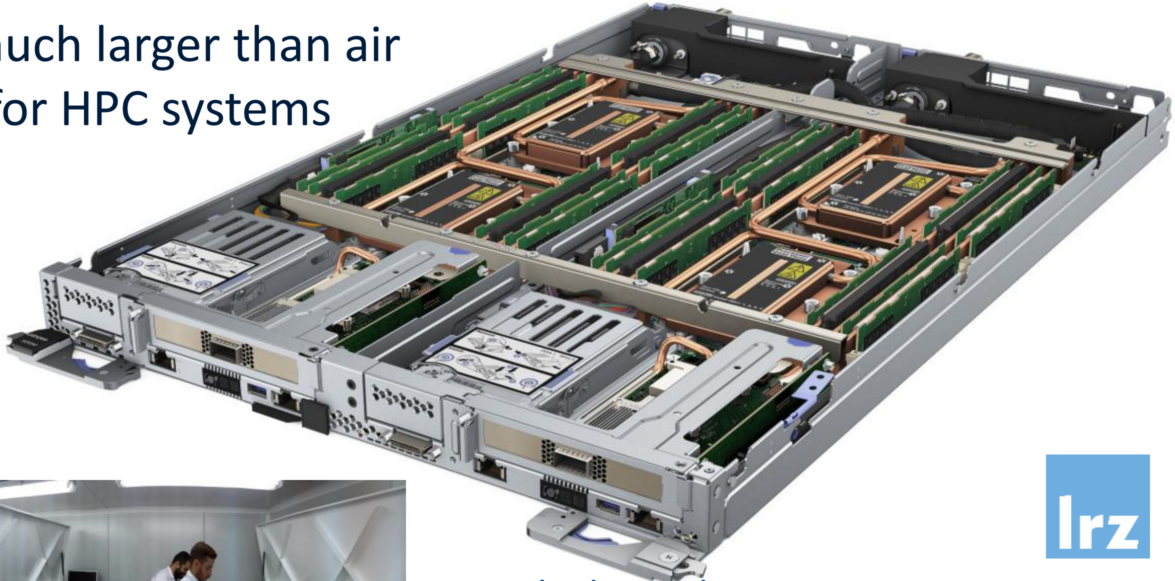


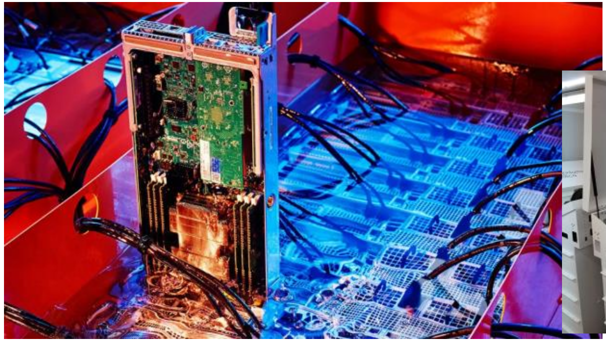
Nikhef 2PA LCO2 cooling setup. Image from Bart Verlaat, Auke-Pieter Colijn *CO2 Cooling Developments for HEP Detectors* https://doi.org/10.22323/1.095.0031

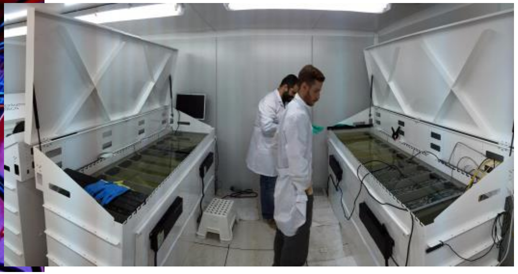
Getting the heat out in liquid form, maybe?

- Heat capacity of liquid is much larger than air
- by now (almost) standard for HPC systems

- immersive systems look cool, but are a bit hard on maintenance







Strongly depends on systems engineering: when water inlet temperature can be >40 degC, you have almost always free cooling

Image source dual-board system: Lenovo, ThinkSystem SD650
immersive cooling image https://hypertec.com/blog/sustainable-emerging-tech-liquid-immersion-cooling/, PIC T1 centre, Barcelona, ES

Or scale *inside* one system

- ‘trivial’ step-up is to do multiple sockets in one system
  2-socket, sometimes 4 socket on a motherboard

- to make it appear as a single shared memory system,
  cache coherency is required between the CPUs

- useful for tightly coupled parallel applications
  (weather forecasting, fluid dynamics, climate), but
  not needed for ‘trivially parallel’ high throughput needs

- depending on architecture cache coherency
  kills single-thread performance (although AMD did lot better here than Intel)

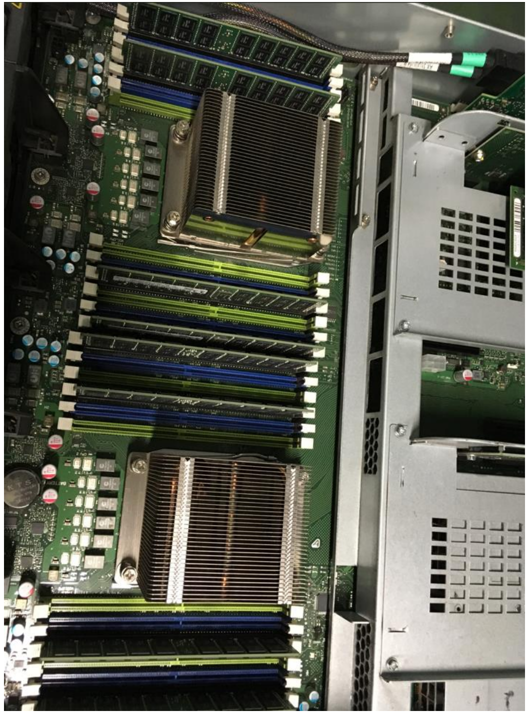


Image: dual-socket Fujitsu system at the Xenon experiment site, 2019. source: Tristan Suerink, Nikhef

CPU design changes may fit application, or not

AMD EPYC effective for applications like WLCG:

- Naples → Rome added shared memory die
- links all cores directly to memory

Rome-Milan improvement?

- shared L3 cache benefits tightly coupled HPC, but not WLCG 'HTC'

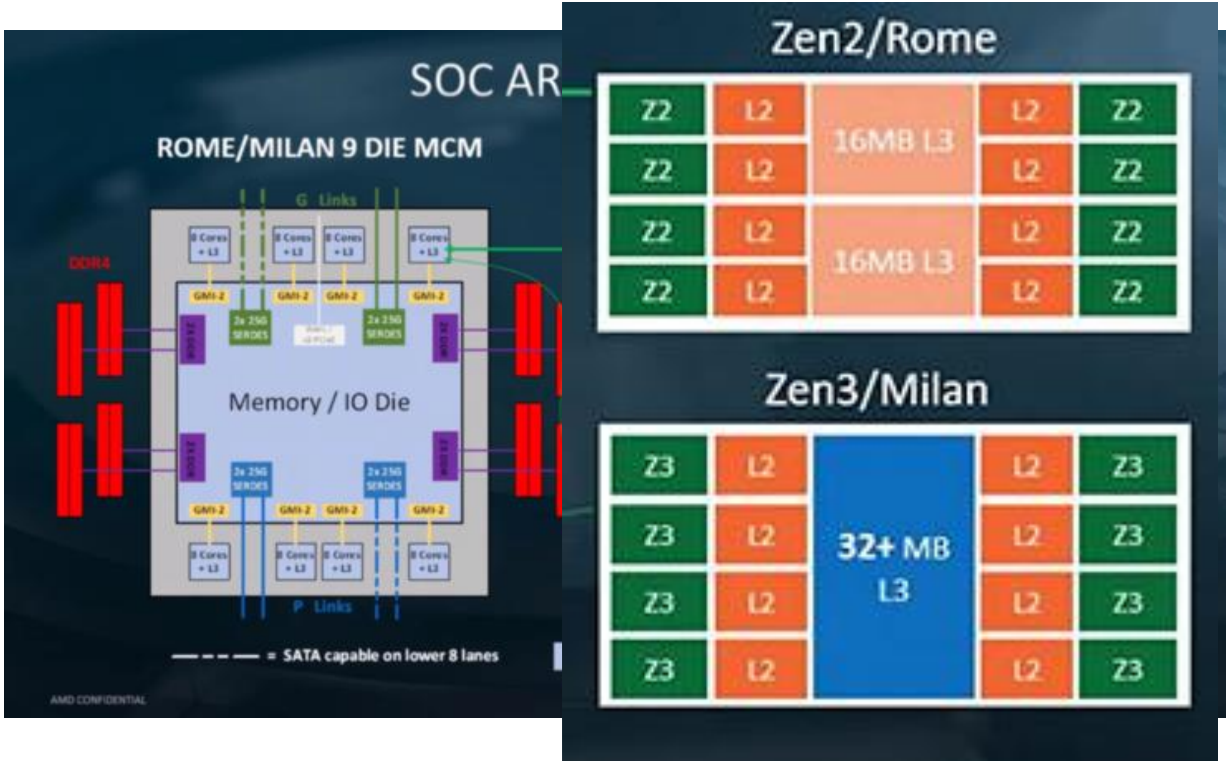
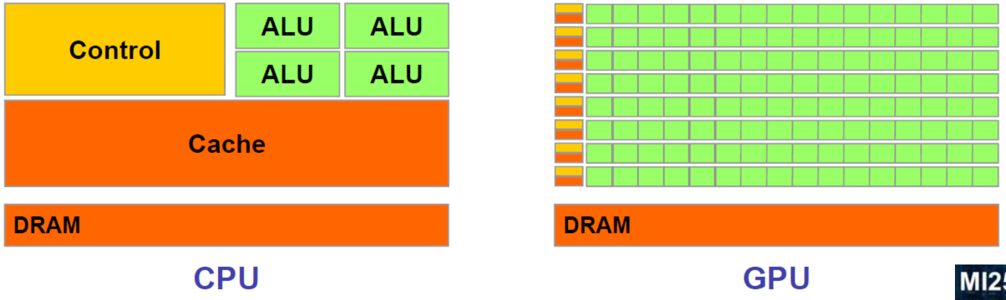


Image source: AMD, retrieved from https://m.hexus.net/tech/news/cpu/135479-amd-shares-details-zen-3-zen-4-architectures/

Accelerators – general purpose GPUs



- but co-processing comes at a cost of moving data to and from the GPU
- often faster to keep computing and do selection & conditionals later
- computation speed heavily depends on precision (even 4-bit precision is used)
- quite power hungry!

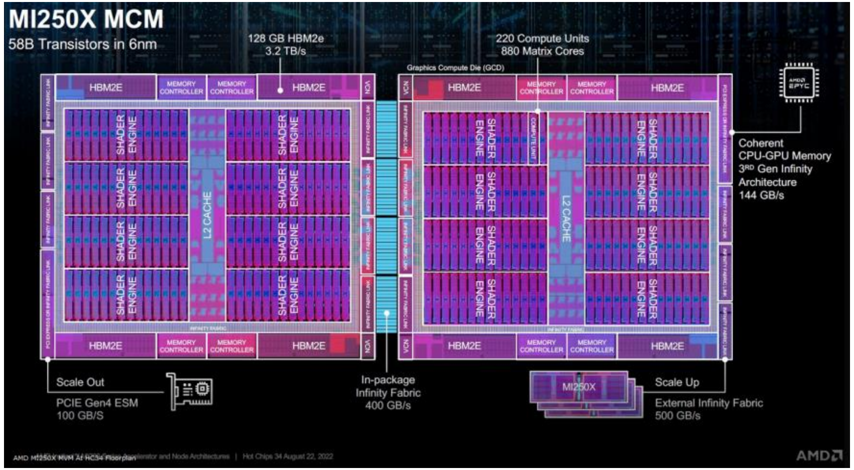


Image: ‘Massively Parallel Computing with CUDA’, Antonino Tumeo Politecnico di Milano, https://www.ogf.org/OGF25/materials/1605/CUDA_Programming.pdf
Floorplan image of die: AMD MI250 GPU, slide source: AMD

If large-scale IT does not quite fit … ahum …

SuperMicro (branded as 'Lambda Blade')
4U chassis, supporting 10 consumer-grade GPUs …
… with a bump

Image source: https://lambdalabs.com/products/blade

Scaling up – beyond one lone motherboard

Physical farms: selecting the 'worker nodes'

For HTC applications
– like WLCG, SKA, WeNMR – typically

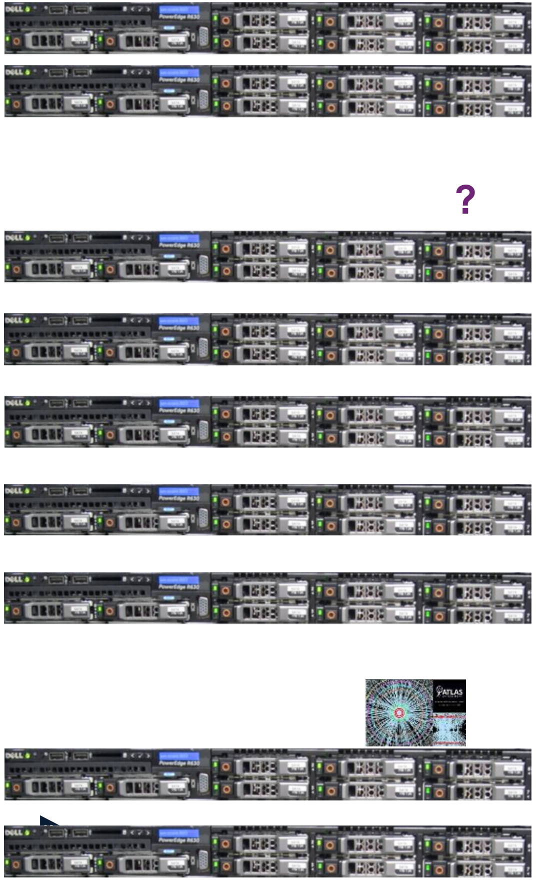
- **balanced features for node throughput**
  (CPU, storage, memory bandwidth, network)

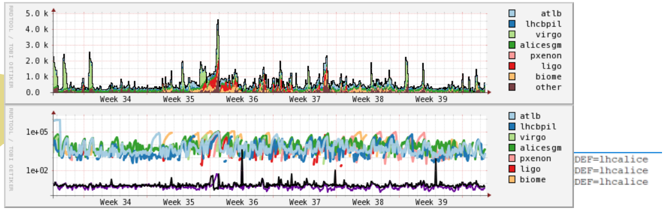
- **single-socket** multicore systems are fine, typical: 64-128 cores per system
- **network**: 2x25Gbps
  (+ 'out of band' management like IPMI)
- **memory**: 8 GiB/core
- **local disk**: 4TB NVME PCIe Gen4 x4

\+ space (physical + power) to add **GPU**

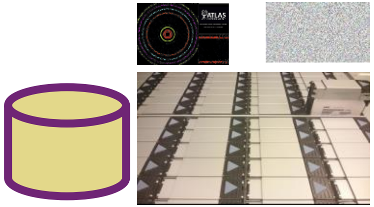
Image: Cluster 'Lotenfeest' at the Nikhef NDPF, acquired March 2020. Lenovo SR655 with AMD EPYC 7702P 64-Core single-socket

WLCG computing – conveniently parallel






```
GROUPCFG[auger]     FSTARGET=3    PRIORITY=200    MAXPROC=500    QDEF=augerbig
GROUPCFG[augsgm]    FSTARGET=1    PRIORITY=300    MAXPROC=2      QDEF=augerbig
QOSCFG[augerbig]    FSTARGET=3

# if these are queued, they will generally be of highest priority.
# limit their MAXIJOBs ... we really want two non-ATLAS VOs to be
# of rank higher than ATLAS before we drain the multicore pool.

GROUPCFG[virgo]     FSTARGET=25    PRIORITY=200    MAXPROC=2700  MAXIJOB=10 QDEF
=biggrid
GROUPCFG[ligo]      FSTARGET=23    PRIORITY=200    MAXPROC=2700  MAXIJOB=10 QDEF
=biggrid

# local groups

GROUPCFG[atlas]     FSTARGET=10   PRIORITY=200    MAXPROC=2200  QDEF=niklocal
```

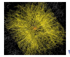
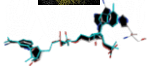


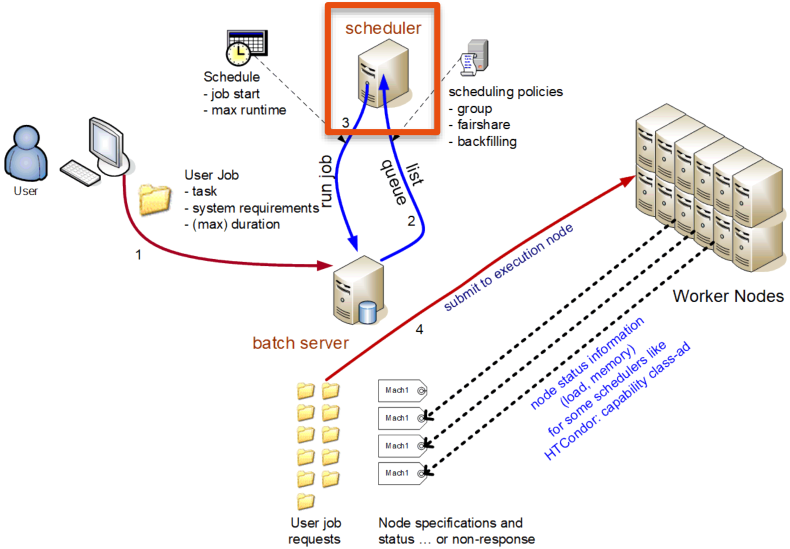
- ‘like milking cows’ (if you feed them lots of power first)
- parallel access to data comes at a cost of high IOPS

Batch system platform

Many things are *conveniently parallel*

- HEP events & simulation
- ligand matching
- structural biochemistry
- …

challenge not in parallelism itself

- we have had HPC systems for ages

but

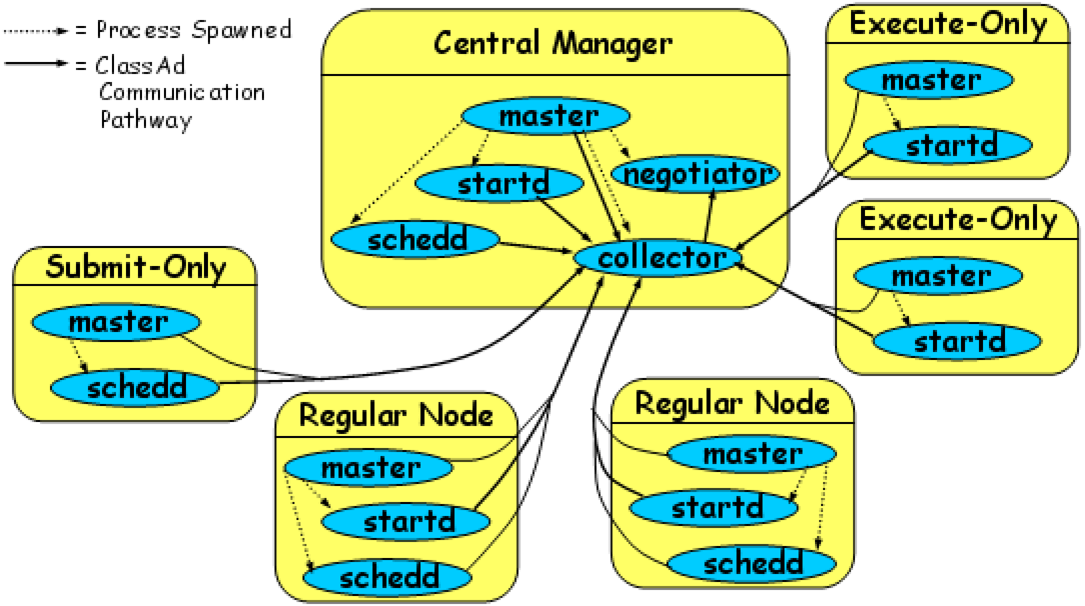
- large numbers of single-core jobs
- heterogeneous workloads
  sharing the same set of worker nodes
- computing with concurrent data access

```
korf.nikhef.nl:
                                                                    Req'd    Req'd       Elap
Job ID                  Username    Queue    NDS   TSK    Memory   Time    S   Time
----------------------- ----------- -------- ----- ------ ------ --------- - ---------
33134895.korf.nikhef.n  lhcbpi08    lhcb         1      1  5120m  41:59:57 R  37:46:21   wn-choc-023
33134901.korf.nikhef.n  lhcbpi08    lhcb         1      1  5120m  41:59:57 R  40:04:09   wn-smrt-128
33134908.korf.nikhef.n  lhcbpi08    lhcb         1      1  5120m  41:59:57 R  37:14:29   wn-choc-030
33134917.korf.nikhef.n  lhcbpi08    lhcb         1      1  5120m  41:59:57 R  14:23:42   wn-smrt-072
33135197.korf.nikhef.n  atlb019     atlasmc      1      4  16040 208:00:00 R 183:02:04   wn-mars-018+
wn-mars-018+wn-mars-018+wn-mars-018
33135883.korf.nikhef.n  atlb019     atlasmc      1      4  16040 208:00:00 R 166:44:22   wn-mars-018+
wn-mars-018+wn-mars-018+wn-mars-018
33142633.korf.nikhef.n  lhcbpi08    lhcb         1      1  5120m  41:59:57 R  37:30:47   wn-mars-043
33149106.korf.nikhef.n  lhcbpi08    lhcb         1      1  5120m  41:59:57 R  10:23:30   wn-car-027
33149132.korf.nikhef.n  lhcbpi08    lhcb         1      1  5120m  41:59:57 R  32:36:49   wn-mars-057
33149220.korf.nikhef.n  lhcbpi08    lhcb         1      1  5120m  41:59:57 R  32:50:19   wn-choc-044
33151669.korf.nikhef.n  lhcbpi08    lhcb         1      1  5120m  41:59:57 R  09:49:53   wn-choc-009
33152704.korf.nikhef.n  atlb019     atlasmc      1      4  16040 208:00:00 R 128:39:13   wn-mars-018+
wn-mars-018+wn-mars-018+wn-mars-018
```

Scalable submission: HTCondor

Matchmaking based on 'ClassAds'

- both jobs and machines advertise their requirements and capabilities in 'classified advertisements'
- Matchmaking done by the negotiator
  execution nodes mostly autonomous



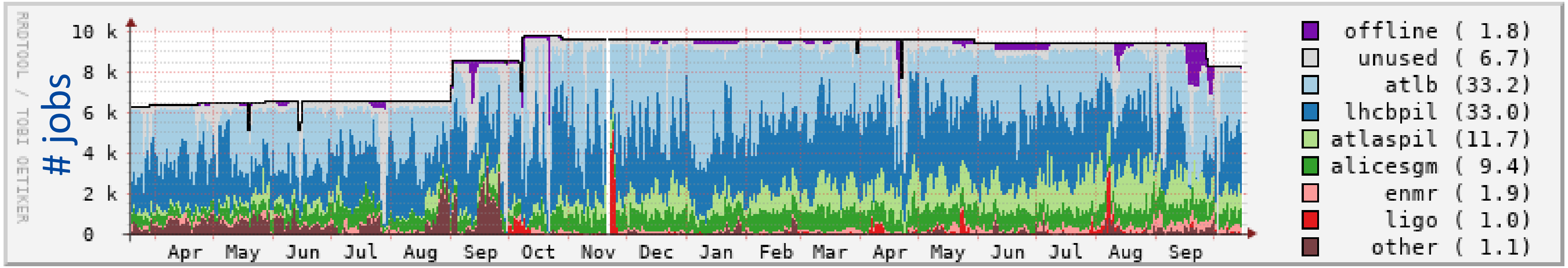
helps for scalability and resilience

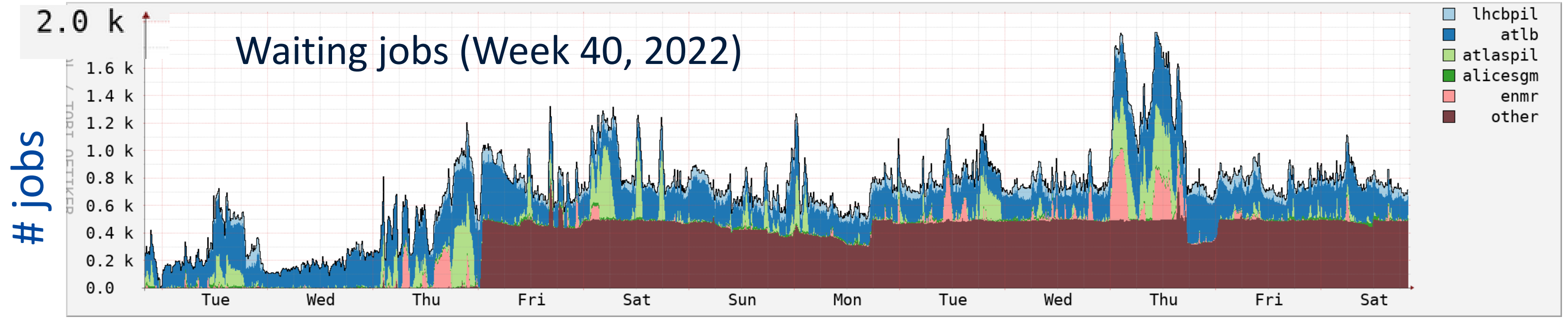
HTCondor, Miron Livny et al, UWMadison; https://research.cs.wisc.edu/htcondor/CondorWeek2008/condor_presentations/desmet_admin_tutorial/

NDPF 'WLCG and Dutch National Infra' cluster

Running jobs:

period: March 2021 .. October 2022



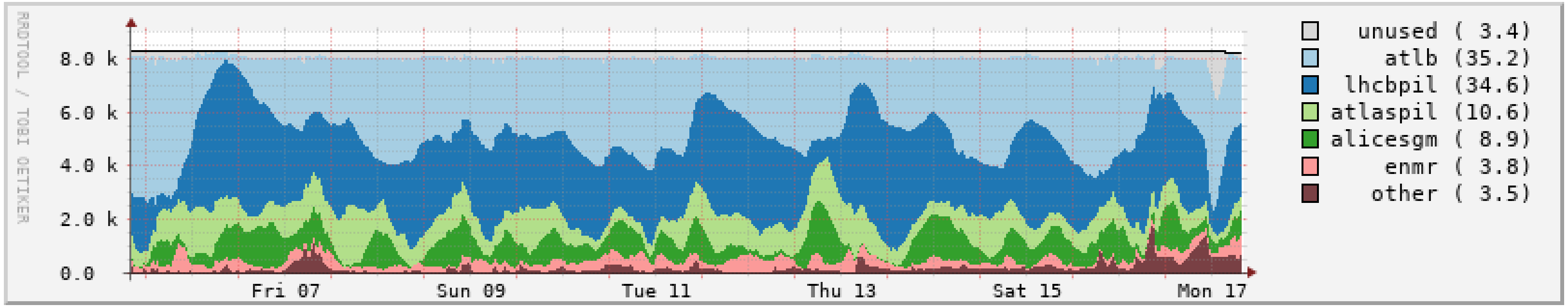


drainage event on Sept 27 are nodes being moved to the LIGO-VIRGO specific cluster; Source: NDPF Statistics overview, https://www.nikhef.nl/pdp/doc/stats/
'other' waiting jobs are almost all for the Auger experiment - GRISview images: Jeff Templon for NDPF and STBC

Estimated Response Time (and predicting it)

- ‘Fair share’ – distributing load over time in a ‘continuous job supply’ system

jobs



wait time (s)

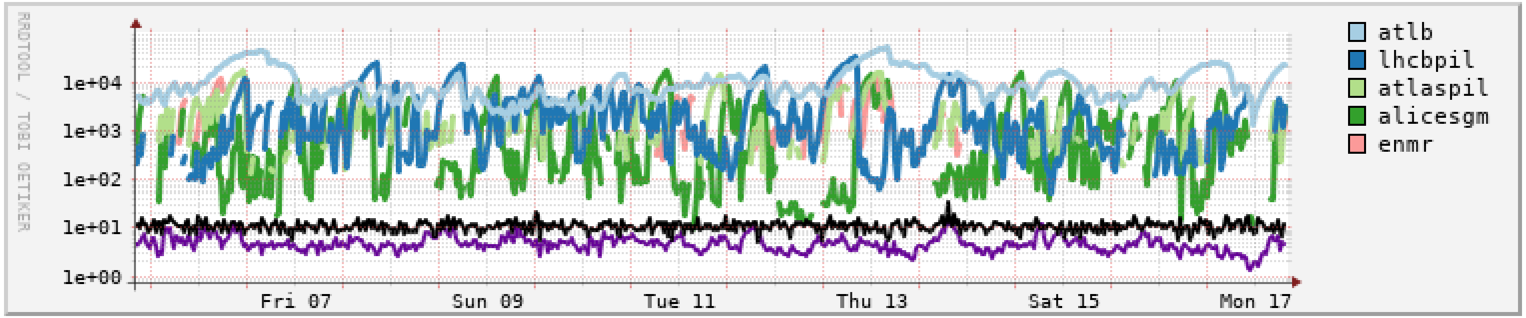


Image: Nikhef NDPF DNI “Grid” cluster. Period: October 6-17, 2022; top-5 communities; GRISview images: Jeff Templon
For work on run time prediction in high-occupancy clusters , see Hui Li *Workload characterization, modeling, and prediction …* https://hdl.handle.net/1887/12574

For occupancy, intended target audience makes a difference

For organized experiment-wide analysis, planned months in advance in WLCG

- ***predictable* scheduling** is more important (steady flow of results)
- **maximizing efficiency**: resource cost is the limiting factor in (physics) results
- co-scheduling with data (pre-placement) is required
- community-authorization based access to data sources only

For ‘local’ users, e.g. students whose progress tomorrow depends on results *today*

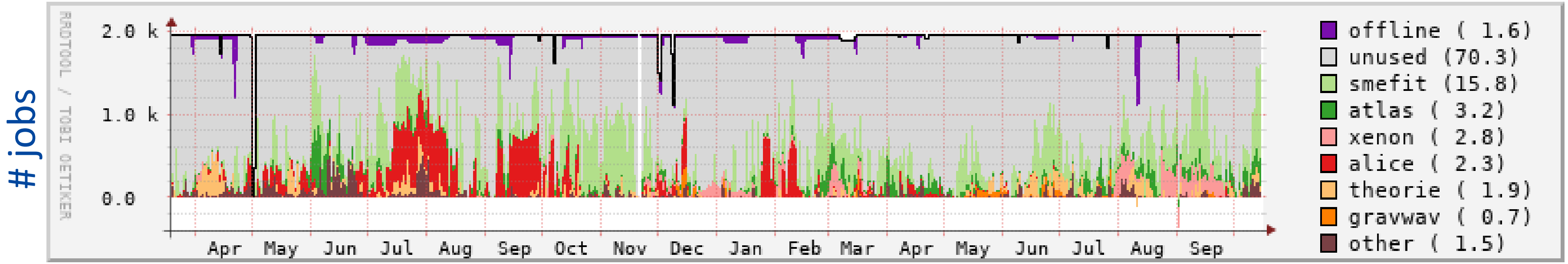
- ***response time*** is more important than efficiency
- fast turn-around/short waiting times
- data access must be parallelism-ready, but is ‘always’ local on-site
- local storage credentials and sharing with desktop and Jupyter environments

so offering two distinct classes of services is (in this case) intentional

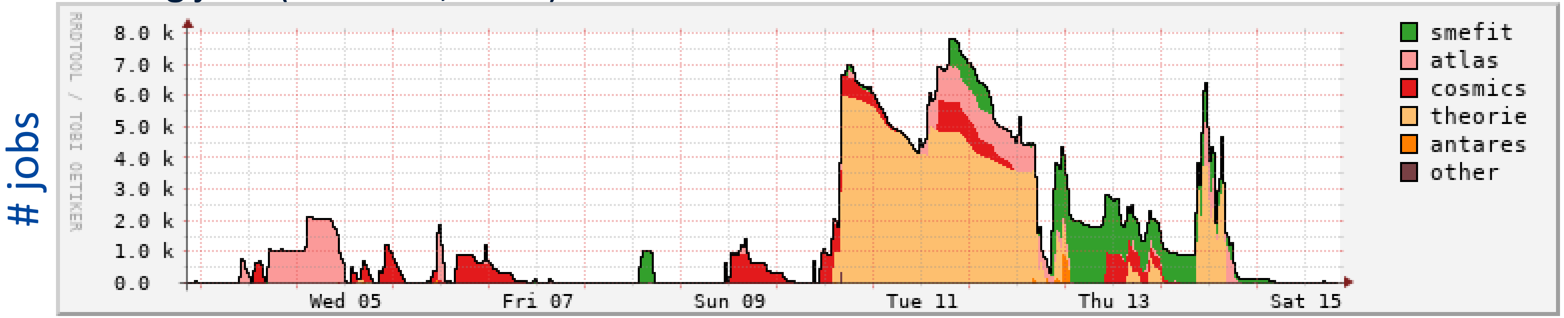
NDPF local analysis cluster 'Stoomboot'

period: March 2021 .. October 2022

Running jobs:



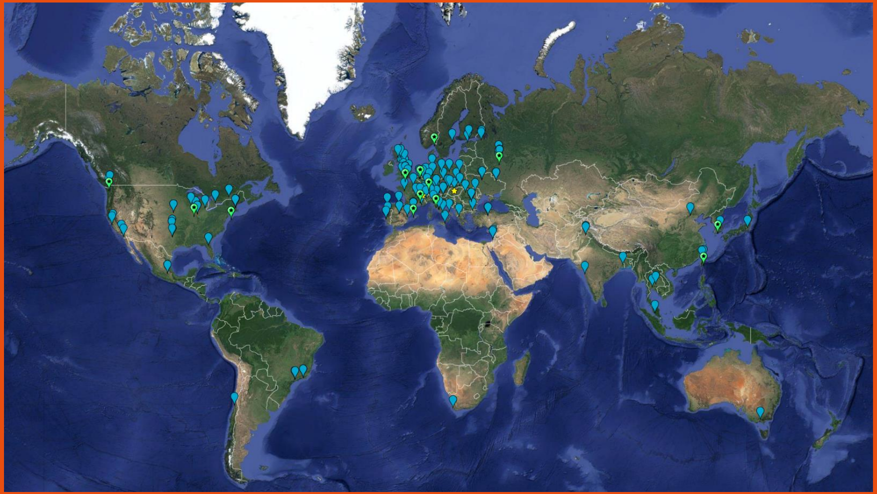
Waiting jobs (Week 40, 2022):



Source: NDPF Statistics overview, https://www.nikhef.nl/pdp/doc/stats/ - GRISview images: Jeff Templon for NDPF and STBC

More of *more than one …*

More than one system
More than one site
More than one user group
More than one organization
More than one …



worldmap: background image google earth, pins indicate WLCG resource centres;

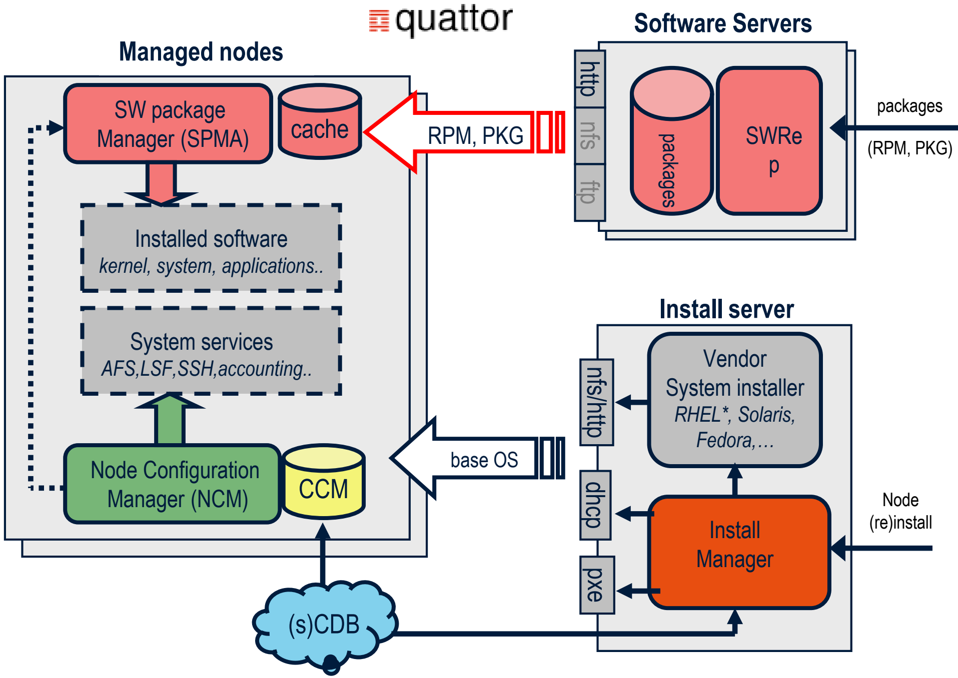
Fancy an interactive console install?

Images: Nikhef Housing H234b NDPF science processing data centre

Managing multiple nodes – *also virtual ones*

Fabric (Configuration) Management

- do you know what is out there?
- update quickly & consistently when vulnerabilities are found?
- versioned repository for rollback?



note that not all tooling scales in itself

- **push**: ansible using ssh logins, or home-brewn scripting
- **pull**: each node runs its own actions, e.g. Quattor, Saltstack, ansible-agent, …

Illustration: German Cancio, CERN, quattor.org, used here as example; see also: ansible.com, saltproject.io, theforeman.org, cfengine.com, puppet.com, …

Scaling 'as a service'

The managed servers usually are not physical

- although there is lots of 'fixed' virtualization of systems, network and (block) storage

When scale, or environment, must be flexible, you get ***software defined infrastructure***

- IaaS: Infrastructure as a Service
- PaaS: Platform as a Service (containers, but also a batch system …)
- SaaS: Software as a Service (like the WeNMR portal)

driven from a configuration management DB

powerful tools, but also easy to get wrong (i.e. having plain-text secrets in the version control system to automate redeployment). And abstractions are *leaky*!

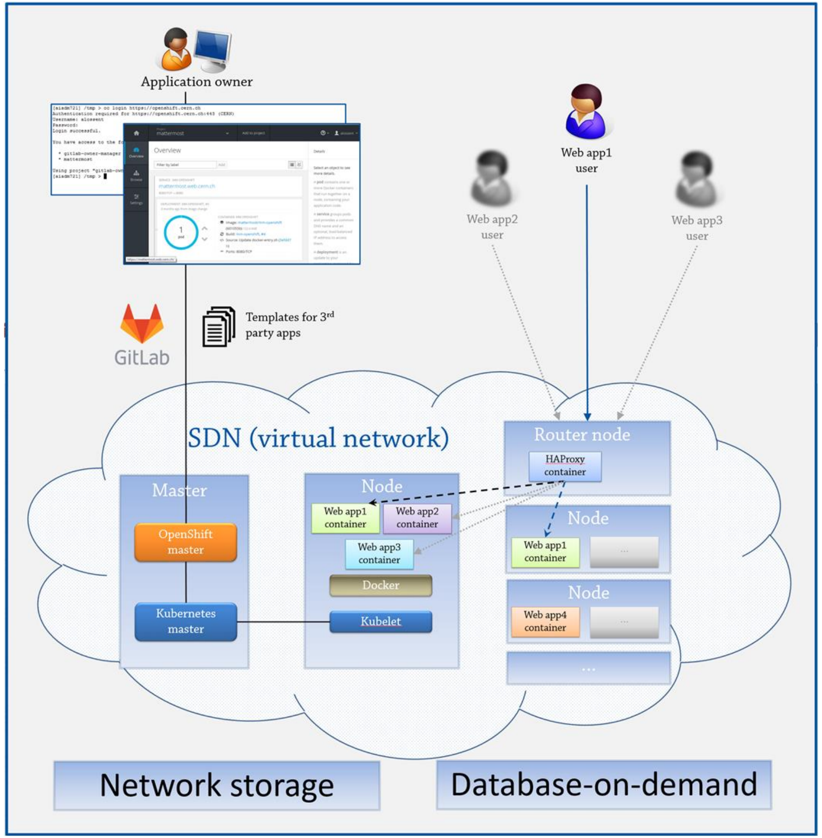
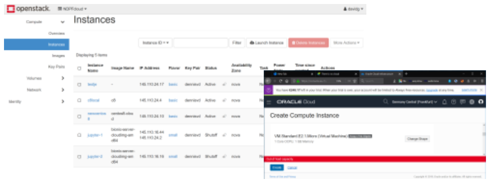
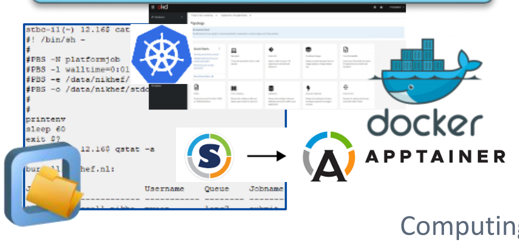


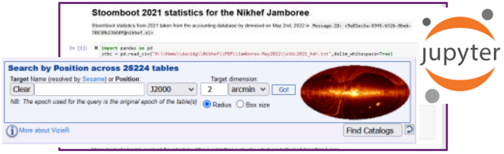
Image from CERN's OpenShift, A Lossent *et al* 2017 *J. Phys.: Conf. Ser.* **898** 082037 https://doi.org/10.1088/1742-6596/898/8/082037

Moving the management boundary

	Infrastructure-as-a-Service	Platform-as-a-Service	Software-as-a-Service
	Application	Application	Application
	Data	Data	Data
Guest	Runtime environment	Runtime environment	Runtime environment
	Middleware	Middleware	Middleware
Hyper	Operating system	Operating system	Operating system
visor	Virtualisation layer	Virtualisation layer	Virtualisation layer
	Physical server	Physical server	Physical server
Host	Storage devices	Storage devices	Storage devices
	Network	Network	Network







Astronomy catalogue: https://vizier.cds.unistra.fr/

IaaS: openstack.com, Oracle OCI; PaaS: dsri.maastrichtuniversity.nl, apptainer.org, cvmfs.readthedocs.io, kubernetes.io, slurm.schedmd.com; SaaS: Jupyter.org

There is NO CLOUD, just other people's computers

Image source: Free Software Foundation Europe - https://fsfe.org/

Brief look at data centres

- ‘tier-1’ … ‘tier-4’ datacenters - increasingly redundant
- all systems are ‘lights out’, since the DC may be miles away
    - remotely controlled, incl. power-on, remote KVM
- small and large in terms of power and cooling capacity
    - Nikhef ~2 MW, Meta Zeewolde would have been 160 MW

- data centre efficiency metric: $PUE = \frac{E_{total}}{E_{IT\_equipment}}$

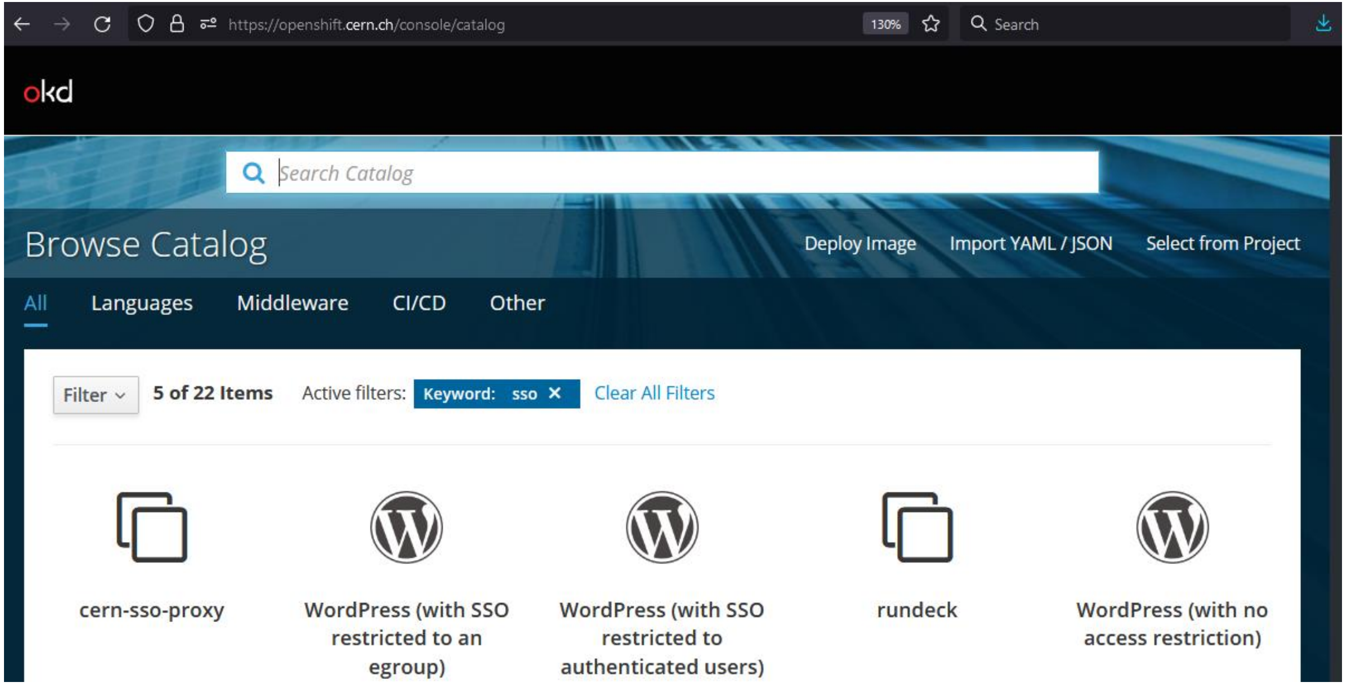
Current Power	Minimum Power	Peak Power	Average Power	Current / Maximum Power	
264 Watt	264 Watt	273 Watt	267 Watt	264	480 Watt

Reducing cost and impact by improving “Power Unit Efficiency” of the data centre:
- airflow engineering and efficient CRACs
- (free) cooling by changing inflow temperature
- Aquifer Thermal Energy Storage (ATES) to buffer heat (and re-use later for homes)

Typical PUEs vary from 1.03 (in Iceland) to 1.2 for ‘good’ datacenters in NL

Data centre tiering: Uptime Institute (Tunner, W.P.; Seader, J.H.; Brill, K.G. Tier Classifications Define Site Infrastructure Performance; White Paper)
Remote systems management: IPMI, RedFish and various vendor proprietary solutions – usually dedicated ‘out-of-band’ network connection, incl. remote KVM

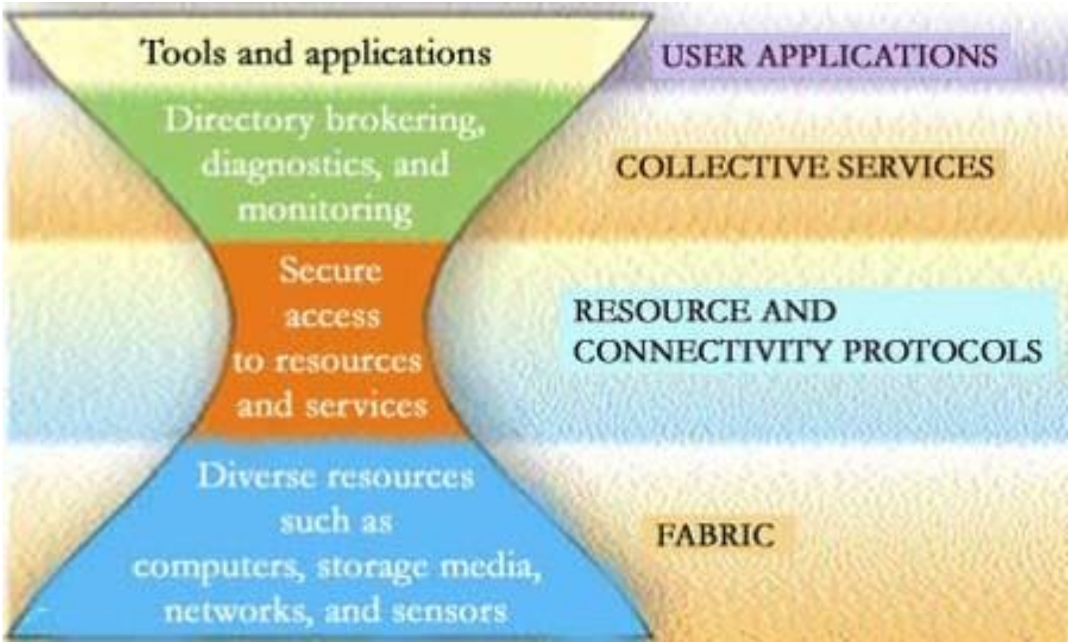
‘Cloudification’ eases systems management …



OpenShift (OKD) system at CERN (accessible for CERN users only);

Common interfaces to the different clouds?

‘protocol hourglass’



hourglass image: Alessio Merlo in The Condor on the Grid: state of art and open issues,

Standard interfaces for compute and data?

‘hourglass’ model kind-of worked for IP, and ~ web with http as common standard

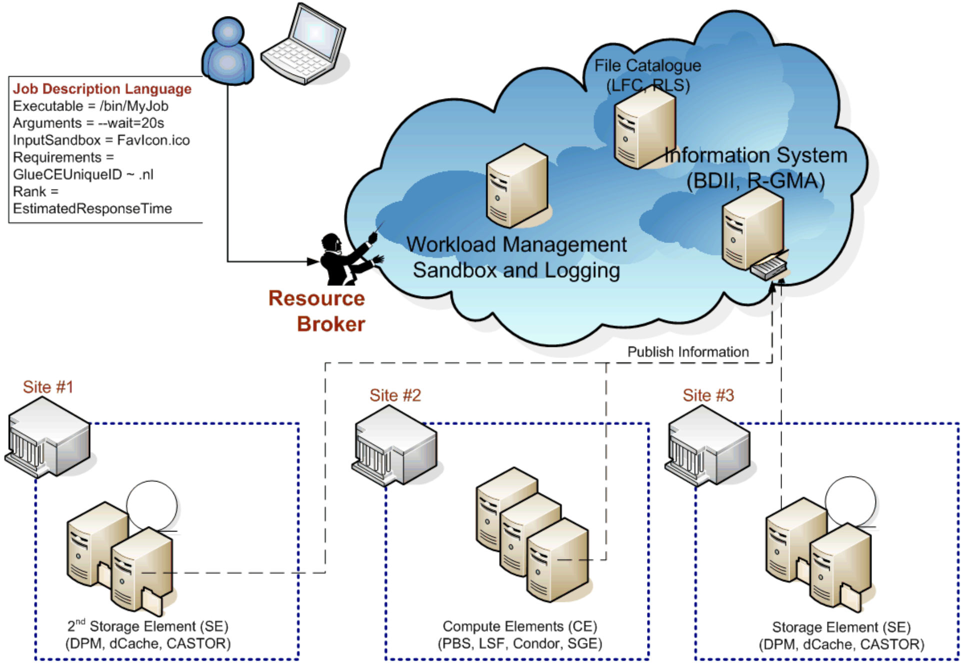
- a very simple stateless interface

protocols for higher-level services never reached this level of global interop

- requirements too complex and stateful
- use cases were usually scoped

slowly changing now but only for similar simple things like on-line object storage

Is distributed computing too bespoke …?



Interoperable cloud? Compare OGF’s OCCI WG GFD.221 (https://www.ogf.org/documents/GFD.221.pdf) with e.g. Amazon S3 API or the OwnCloud CS3 interfaces

DIRAC: spanning heterogeneous resource models

Adding a scheduling layer on top

all sites in WLCG are autonomous – and global standards failed

'any (IT) problem can be solved by adding one layer of indirection'

DIRAC is just one example

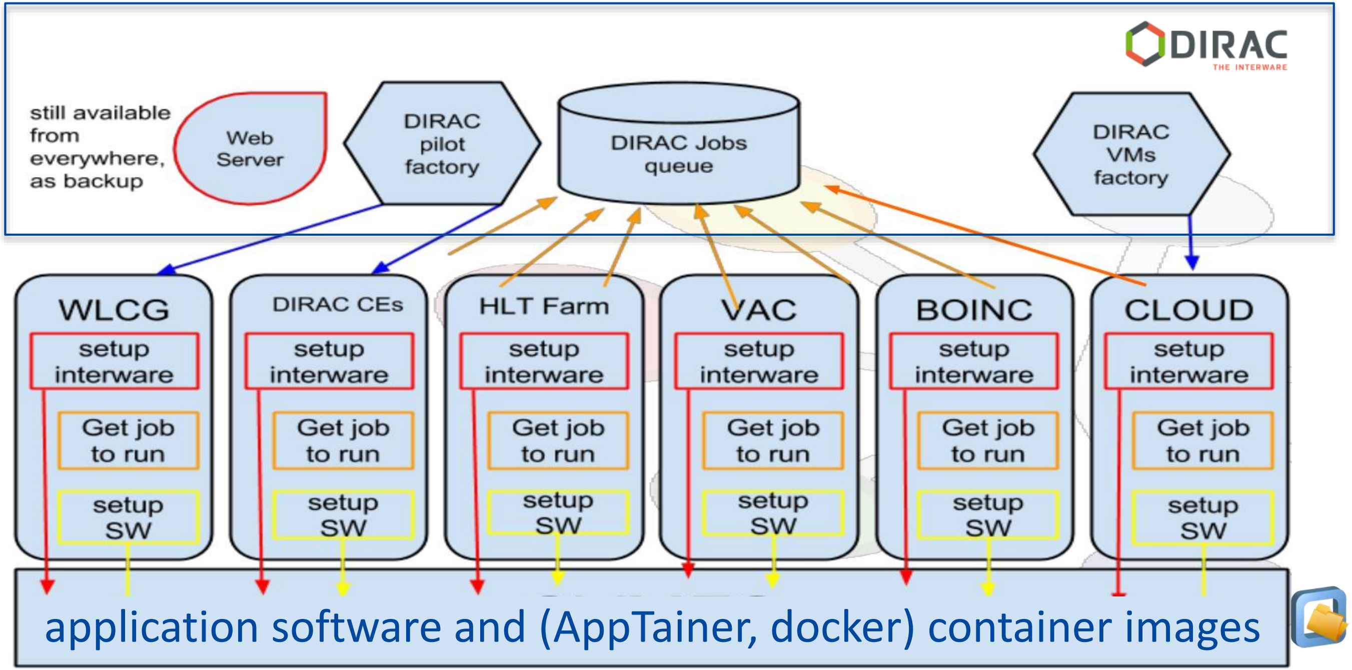
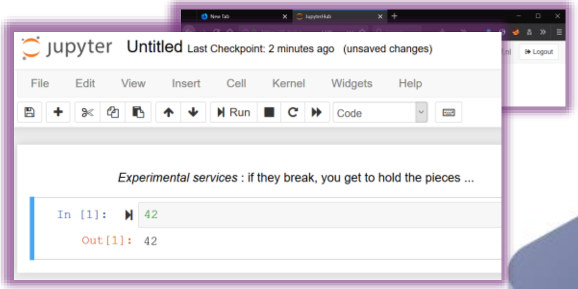
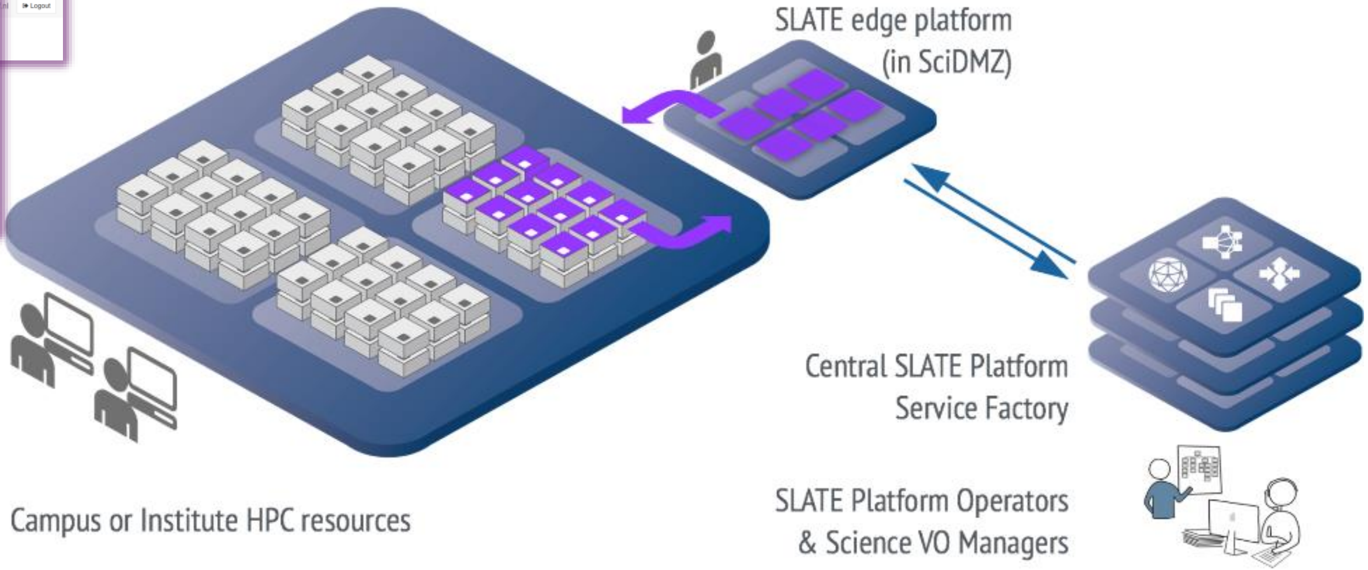


Image: DIRAC project, A. Tsaregorodtsev *et al.* CPPM Marseille, from https://dirac.readthedocs.io/ ; CVMFS (CERN VM File System) is a common software distribution platform using distributed signed data objects in a cached hierarchy using CDN techniques, see https://cernvm.cern.ch/fs/

An overlay network of containers

Nobody wants a cloud per-se … what folk want is a solution …

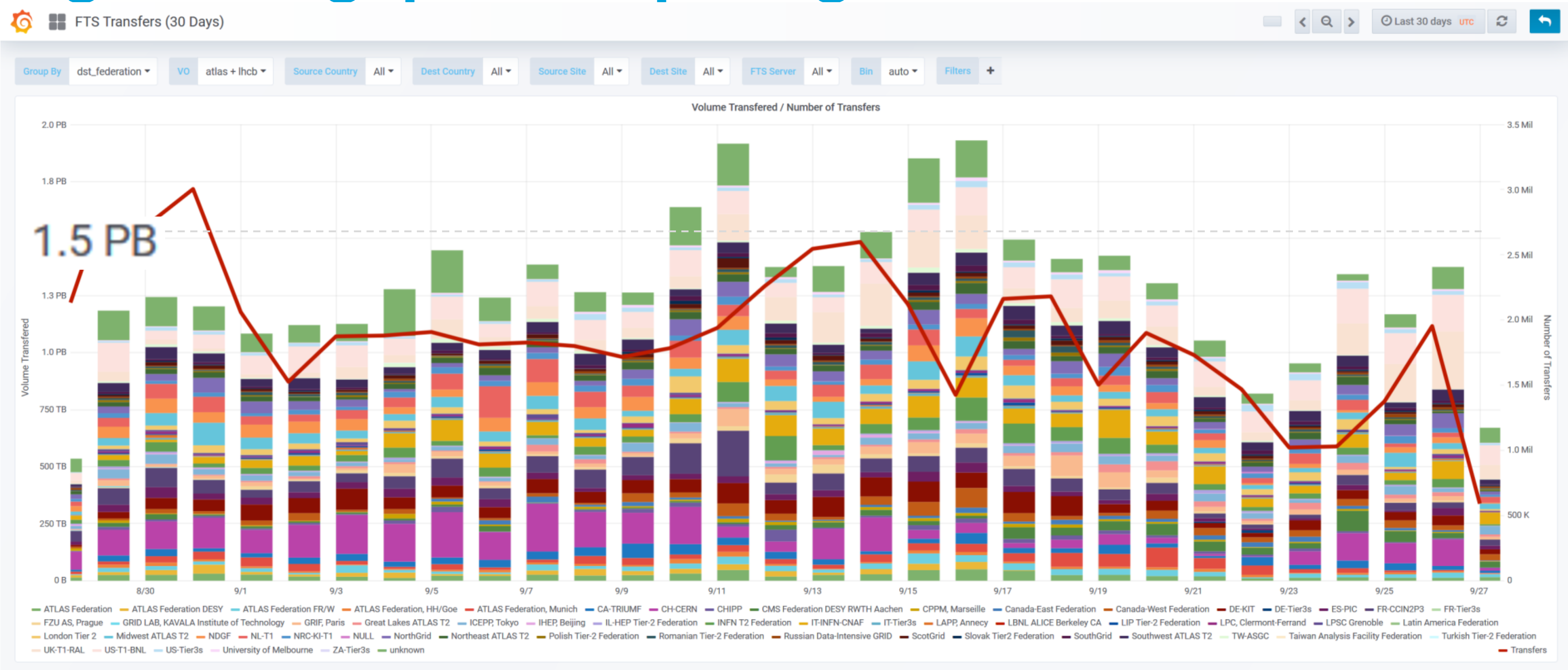




‘alien containers’ HPC integration - container computing, using curated application images

Image sources: NDPF JupyterHub service “Callysto”; SLATE: Service Layer At The Edge – Rob Gartner (UChicago), Shawn KcMee (UMich) *et al.* – slateci.io

High throughput computing is also about data



source: https://monit-grafana.cern.ch/d/000000420/fts-transfers-30-day ; data: November 2020 ; CERN FTS instance WLCG: daily transfer volume ATLAS+LHCb

Can storage support your parallel processing

Basic storage properties

- throughput
- IOPS – I/O Operations per Second
- seek-time

but not many storage systems support *concurrent parallel access* by many clients

- both data **and** (file system or index) meta-data must be scalably distributed
- typically sacrifice either instant consistency, or (POSIX) semantics, (or scalability) in a distributed storage system

Common commercial solutions: GPFS, (and still: CXFS), … but also NetApp, HDS, Dell-EMC, &c
Common open source: gluster, dCache, CephFS, Lustre, …

And storage is usually *tiered* – fast local → online (spinning) disk → near-line (tape)

Example: client-side managed GlusterFS

- scalable through independence of both clients and servers
- design is stateless: file system meta-data kept in each server’s file system
- data itself can be replicated and protected but … inconsistencies in metadata linger around the corner in case of client failures (e.g. batch system worker nodes)

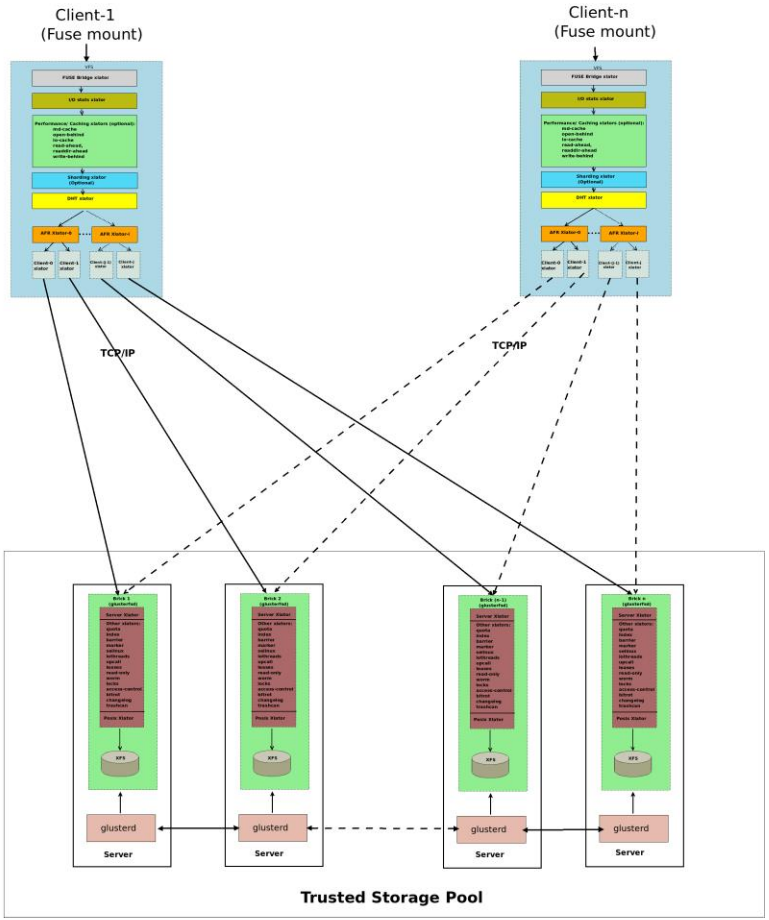
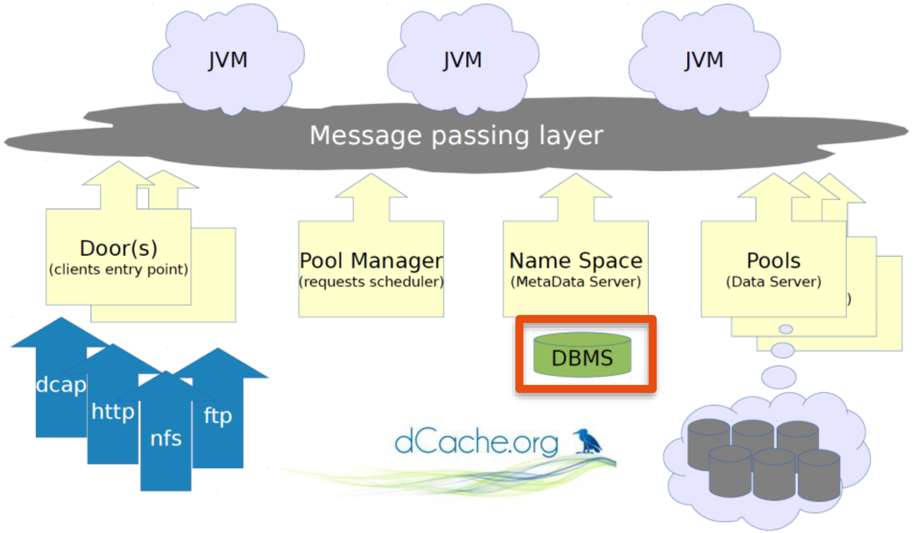
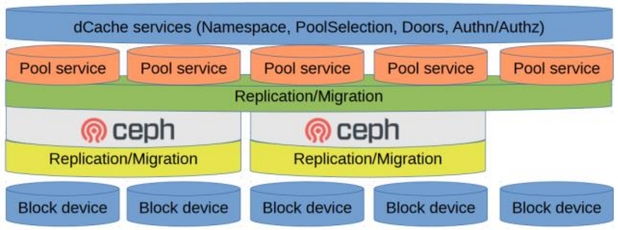


Image source Gluster community: https://docs.gluster.org/en/main/Quick-Start-Guide/Architecture/

Example: server-coherent distribution – dCache

- separate client entry points, storage access scheduling, filesystem meta-data (namespaces), and storage
- message layer for eventual consistency
- redirect-based access
  - doors and pools usually on all nodes
  - now also feature of standard NFSv4.1





Images: Tigran Mkrtchyan (DESY, dCache.org), *dCache on steroids - delegated storage solutions*, ISGC 2016, https://dcache.org/manuals/publications.shtml

dCache: wide area distribution

- can be widely (long latency) distributed
  - Nordic Data Grid Facility: Sweden is quite long (~16ms RTT), and Ljubljana to Umeå is ~30ms RTT (~ 2900km)
- redirect-then-access model limits interactions with any single node across a long-distance links
- at 'cost' of POSIX features like *atime* or concurrent write
  - most distributed applications don't need these anyway
  - but indeed it's not a good backing store for databases ☺

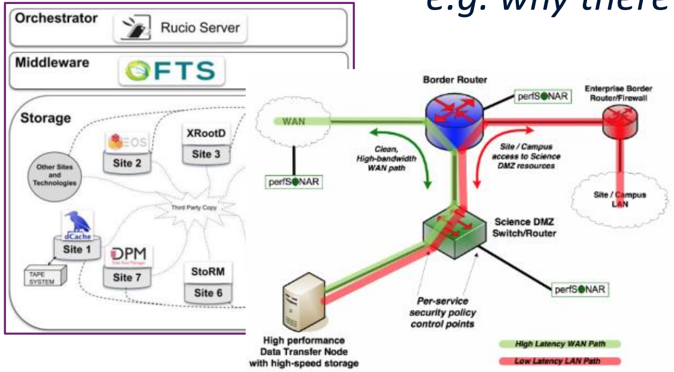
The NDGF dCache instance spans datacentres across Scandinavia and Slovenia, but is administered and used as a single instance.

Image NDGF instance: Jürgen Starek et al. (dCache team) at https://www.dcache.org/manuals/dCache-Whitepaper.pdf; https://dcache.org/manuals/Book-8.2

Structure of application data placement impacts storage (hardware) systems design

pre-staging all data locally supports latency hiding,
posix-style access with lseek(2), and fast local ‘$TMPDIR’

e.g. why there are Data Transfer Nodes (DTNs) in the ‘Science DMZ’ concept



but, nowadays, pre-staging started coming at a cost, when using **SSDs** as local ‘scratch’ area … because of their hardware characteristic ‘endurance’

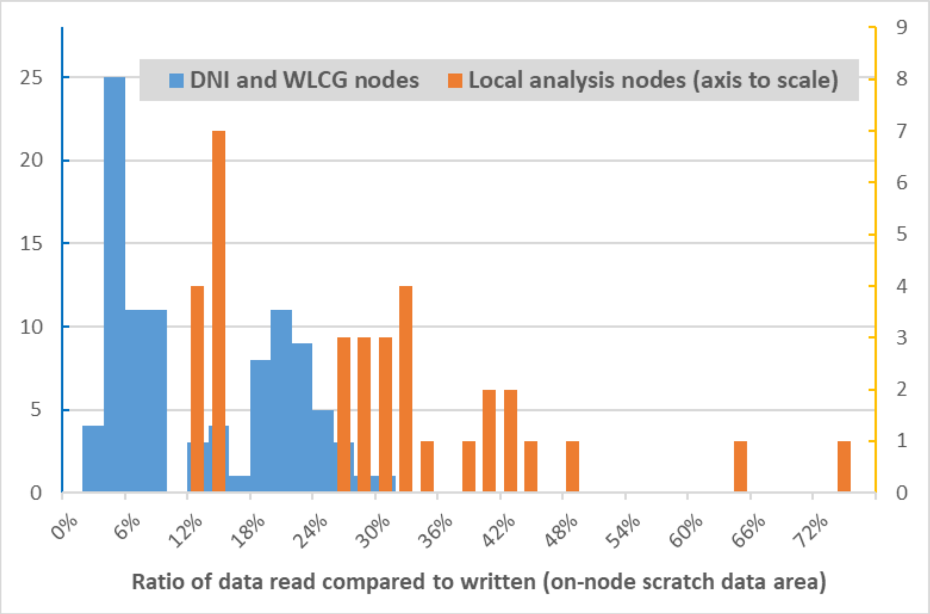
Photo HGST nVME from: Dmitry Nosachev on Wikimedia Commons CC-BY-SA; Image Science DMZ and Data Transfer Nodes: ESnet fasterdata.es.net

Especially with *WORN* storage: Write Once Read Never

Frequency distribution of **read-back vs. write** volume, observed on local scratch for NDPF execution nodes for ***outside ('grid') access (blue) vs local access (orange)***

Access pattern is rather different. But why?

- external users pre-stage, because that is built into the frameworks (like DIRAC, Athena), whereas 'local' users streaming data ('dCache NFSv4')
- different types of workload:
  ntuple-data analysis vs (re)processing



Data: NDPF execution nodes, based on SSD SMART data, integrated over total device lifetime; plot shows number of local analysis nodes scaled to DNI-WLCG count; collected using smartctl on 2020-10-28 – in total 97 'DNI' and 34 'STBC' SSDs were used in the analysis

Putting ‘more than one’ thing together

Connecting the bits

The Internet Is Not Enough!

‘Elephant streams in a packet-switched internet’

‘You may have plenty of shovels,
but where to leave the sand?’

- wheelbarrow works fine in your garden
- want to send it to different places?
  Use waggons on a train, or ships
- always from A-to-B?
  A conveyer belt will do much better!

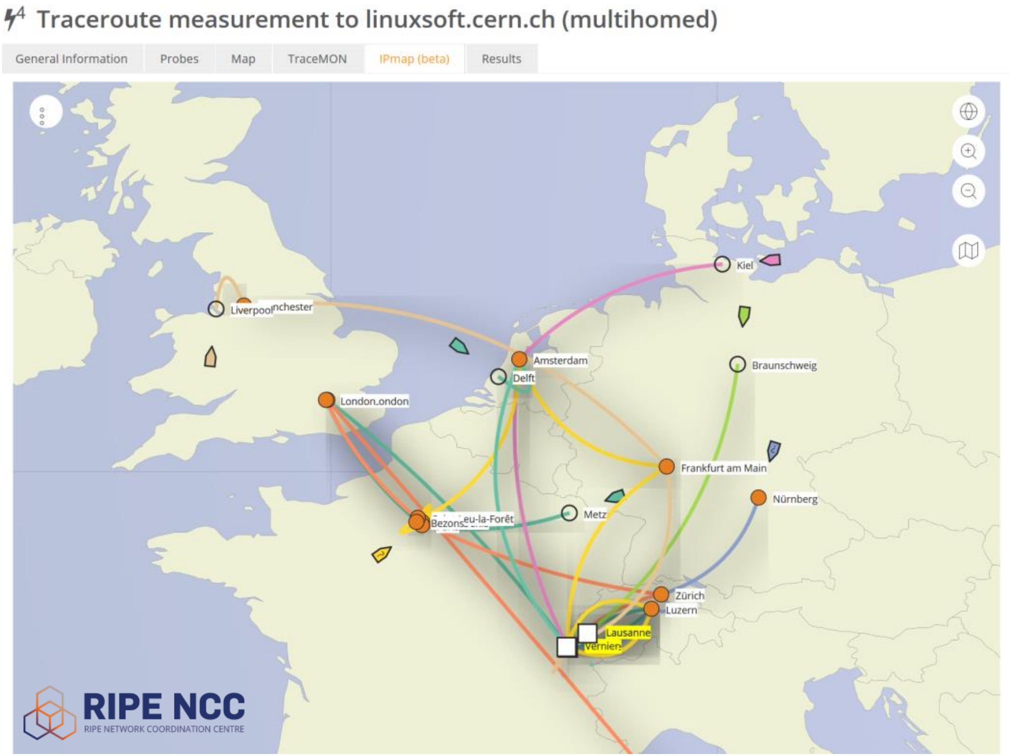
… although you still need
a hole to dump it in …

Image conveyor belt tunnel near Bluntisham, Cambridgeshire by Hugh Venables, CC-BY-SA-4.0 from https://www.geograph.org.uk/photo/4344525

A quick look at internet routing …

network paths
from various places
in Western Europe

towards an IP address at CERN



Data: RIPE NCC Atlas project, TraceMON IPmap, atlas.ripe.net, measurement 9249079

Many paths to Rome … i.e. to your server

- From a home connected to Freedom Internet to *spiegel.nikhef.nl*

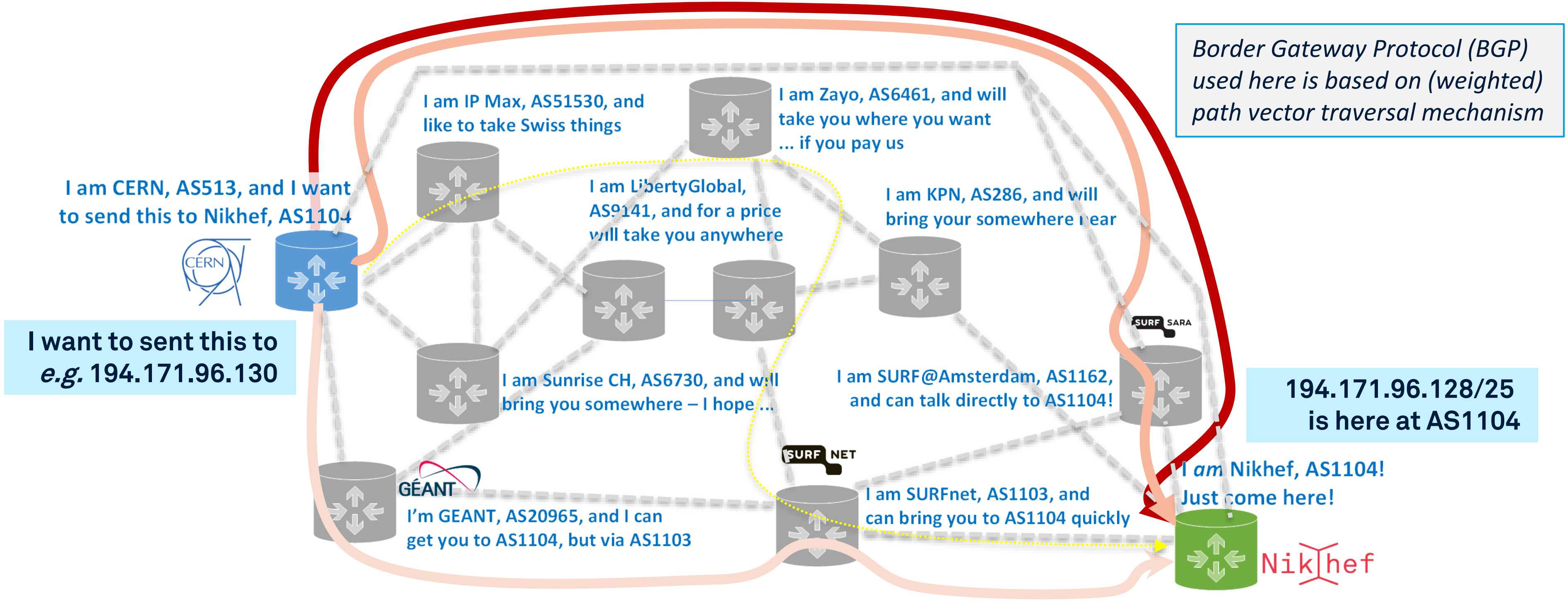
```
[root@kwark ~]# traceroute -6 -A -T gierput.nikhef.nl
traceroute to gierput.nikhef.nl (2a07:8500:120:e010::46), 30 hops max, 80 byte packets
 1  2a10-3781-17b6.connected.by.freedominter.net (2a10:3781:17b6:1:de39:6fff:fe6b:4558) [AS206238]  0.810 ms  1.052 ms  1.330 ms
 2  2a10:3780::234 (2a10:3780::234) [AS206238]  7.460 ms  7.655 ms  7.705 ms
 3  2a10:3780:1::21 (2a10:3780:1::21) [AS206238]  8.868 ms  9.054 ms  9.103 ms
 4  et-0-0-1-1002.core1.fi001.nl.freedomnet.nl (2a10:3780:1::2d) [AS206238]  10.017 ms  9.934 ms  10.263 ms
 5  as1104.frys-ix.net (2001:7f8:10f::450:66) [*]  10.898 ms  11.744 ms  11.797 ms
 6  gierput.nikhef.nl (2a07:8500:120:e010::46) [AS1104]  11.502 ms  7.800 ms  7.357 ms
```

- but from Interparts in Lisse, NH:

```
[root@muis ~]# traceroute -6 -A -I gierput.nikhef.nl
traceroute to gierput.nikhef.nl (2a07:8500:120:e010::46), 30 hops max, 80 byte packets
 1  2a03:e0c0:1002:6601::2 (2a03:e0c0:1002:6601::2) [AS41960]  1.380 ms  1.371 ms  1.369 ms
 2  2a02:690:0:1::b (2a02:690:0:1::b) [AS41960]  1.305 ms  1.312 ms  1.312 ms
 3  et-6-1-0-0.asd002a-jnx-01.surf.net (2001:7f8:1::a500:1103:2) [AS1200]  1.957 ms  2.000 ms  2.052 ms
 4  ae47.asd001b-jnx-01.surf.net (2001:610:e00:2::49c) [AS1103]  2.443 ms  2.505 ms  2.507 ms
 5  irb-4.asd002a-jnx-06.surf.net (2001:610:f00:1120::121) [AS1103]  2.041 ms  2.138 ms  2.138 ms
 6  nikhef-router.customer.surf.net (2001:610:f01:9124::126) [AS1103]  8.977 ms  7.957 ms  7.951 ms
 7  gierput.nikhef.nl (2a07:8500:120:e010::46) [AS1104]  7.922 ms  8.093 ms  8.081 ms
```

AS41960: Interparts; AS1200: AMS-IX route reflector; AS1103: SURFnet; AS1104: Nikhef; AS206238: Freedom Internet – on the FrysIX there is direct L2 peering

Where do internet packets go anyway?



grey-dash lines for illustration only: may not correspond to actual peerings or transit agreements; red lines: the three existing LHCOPN and R&E fall-back routes; yellow: public internet fall-back (least preferred option)

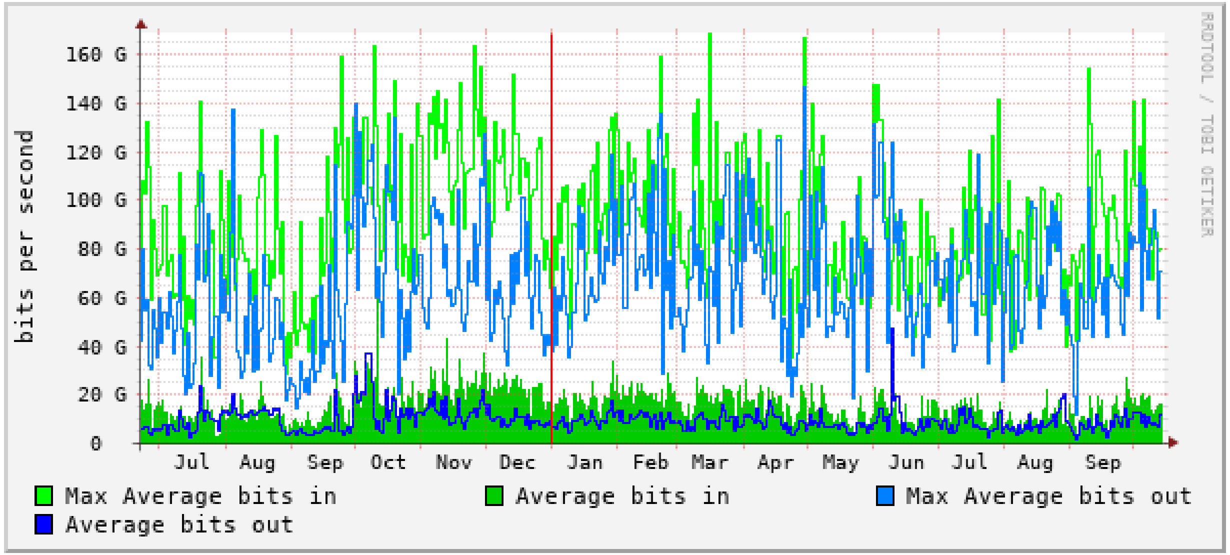
Announcing routes: the Border Gateway Protocol

```
davidg@deelqfx-re0> show route receive-protocol bgp 192.16.166.21 table LHCOPN

LHCOPN.inet.0: 316 destinations, 344 routes (316 active, 0 holddown, 0 hidden)
  Prefix                  Nexthop              MED     Lclpref    AS path
* 109.105.124.0/22        192.16.166.21        10                 513 39590 I
* 117.103.96.0/20         192.16.166.21        10                 513 24167 I
* 128.142.0.0/16          192.16.166.21        10                 513 I
* 130.199.48.0/23         192.16.166.21        10                 513 43 ?
* 130.199.185.0/24        192.16.166.21        10                 513 43 ?
* 130.246.176.0/22        192.16.166.21        10                 513 43475 I
```

```
davidg@deelqfx-re0> show route advertising-protocol bgp 192.16.166.21 table LHCOPN

LHCOPN.inet.0: 316 destinations, 344 routes (316 active, 0 holddown, 0 hidden)
  Prefix                  Nexthop              MED     Lclpref    AS path
* 192.16.186.160/30       Self                                    I
* 194.171.96.128/25       Self                                    I
* 194.171.98.112/29       Self                                    I
```

IPv4 routes advertised from AS513/CERN (for all sites on LHCOPN) to AS1104/Nikhef (top), and the routes announced by AS1104/Nikhef to CERN, on 5 Nov 2022

Typical data traffic to and from the processing cluster



Source: Nikhef cricket graphs period June 2021 – October 2022 – aggregated (research) traffic to external peers from deelqfx – https://cricket.nikhef.nl/

Network is more than just what it says on the tin

More network bandwidth does not mean your *data* gets there faster

- memory requirements (since TCP needs a capability to re-transmit)

- tcp 'slow start'
- congestion control algorithms

TCP throughput calculator

Theoretical network limit

rough estimation: rate < (MSS/RTT)*(C/sqrt(Loss)) [C=1] (based on the Mathis et.al. formula)

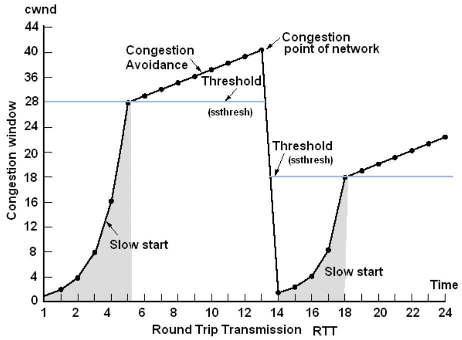
network limit (MSS 9000 byte, RTT: 150.0 ms, Loss: $2.304*10^{-11}$ ($2*10^{-09}$%)) : **100000.00 Mbit/sec.**

Bandwidth-delay Product and buffer size

BDP (100000 Mbit/sec, 150.0 ms) = **1875.00 MByte**

required tcp buffer to reach 100000 Mbps with RTT of 150.0 ms >= **1831054.7 KByte**

maximum throughput with a TCP window of 1831054 KByte and RTT of 150.0 ms <= **100000.00 Mbit/sec.**



Useful sources: https://www.switch.ch/network/tools/tcp_throughput/, https://fasterdata.es.net/

tcp slow-start graphic from Abed et al, *Improvement of TCP Congestion Window over LTE- Advanced Networks* IJoARiC&CE 2012

That viral cat video destroyed it all …

latency AMS-GVA 17 ms
congestion event @20ms:
2 ms of UDP traffic to GVA

- TCP protocol sensitive to packet loss
  - 3 lost packets is enough to trigger this
- different congestion avoidance algorithms exists (~20 by now)
- loss severely impacts links w/large 'bandwidth-delay-product' (BDP)
- NL: ~3 ms, US East: 150ms

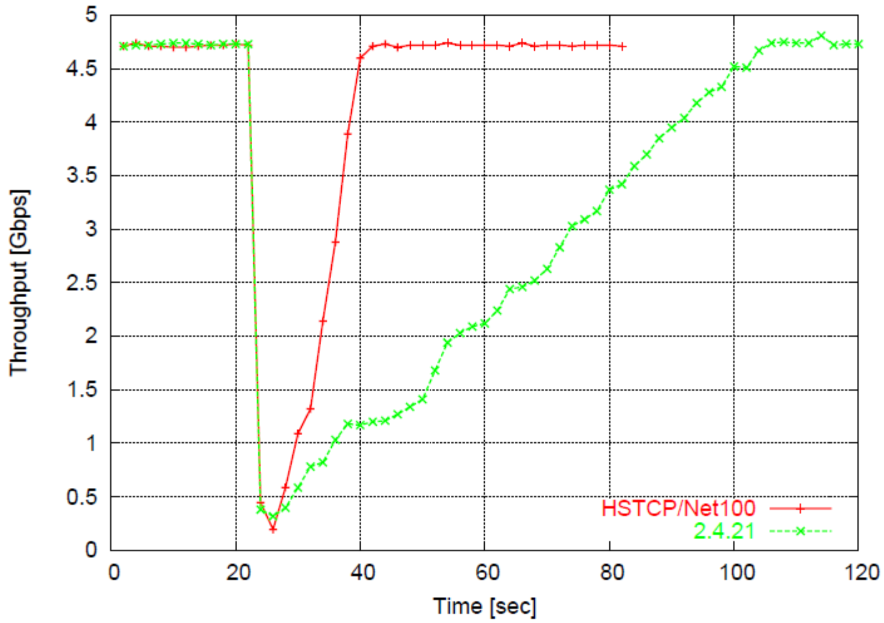


Figure 10: HSTCP versus stock TCP recovery time

source: Catalin Meirosu et al. *Native 10 Gigabit Ethernet experiments over long distances* in FGCS, doi:10.1016/j.future.2004.10.003 – aka. ATL-D-TN-0001

LHCOPN – distributing raw data

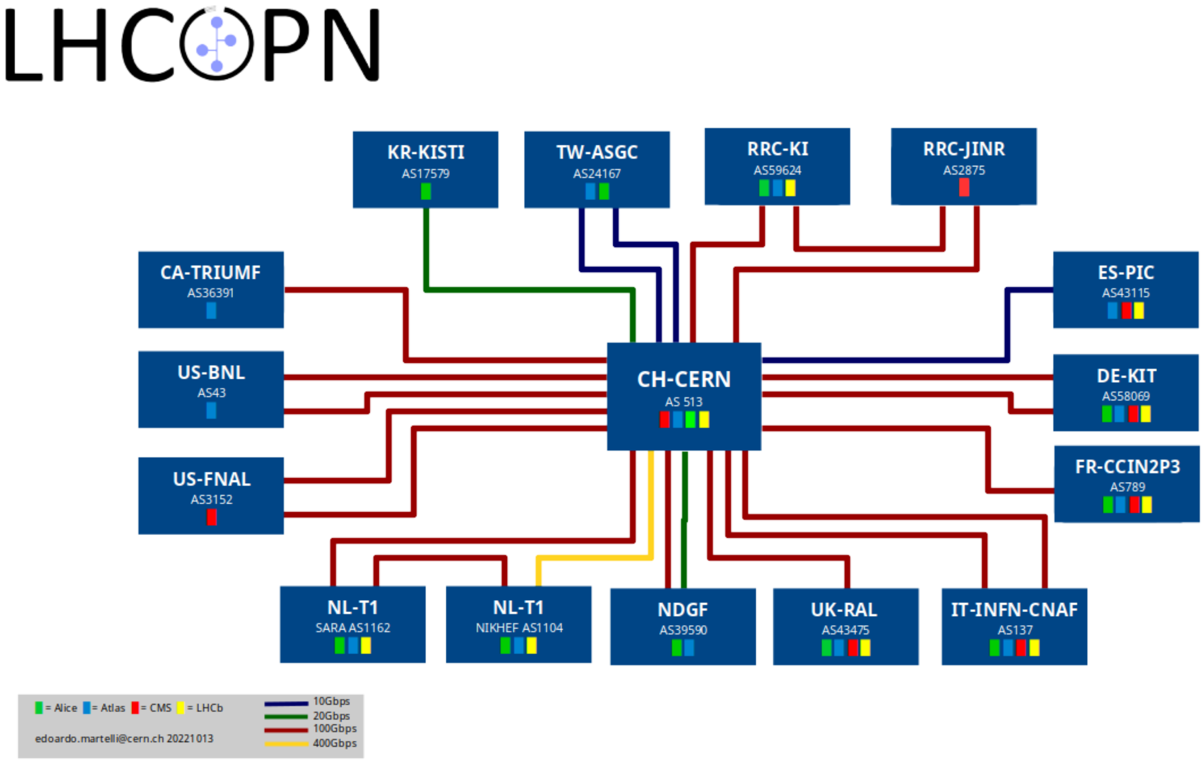
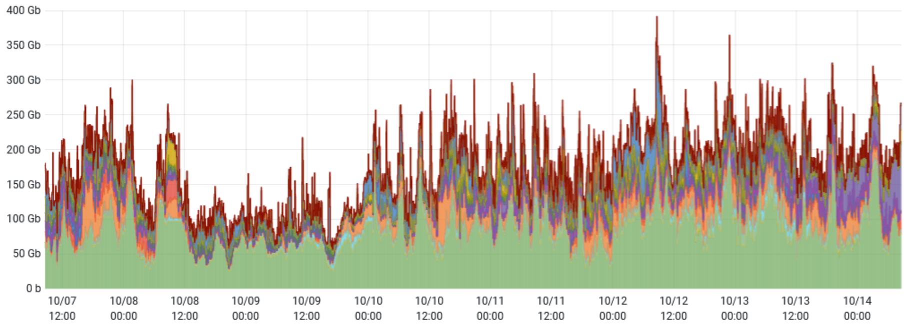


Image source: Edoardo Martelli, CERN, https://lhcopn.web.cern.ch/

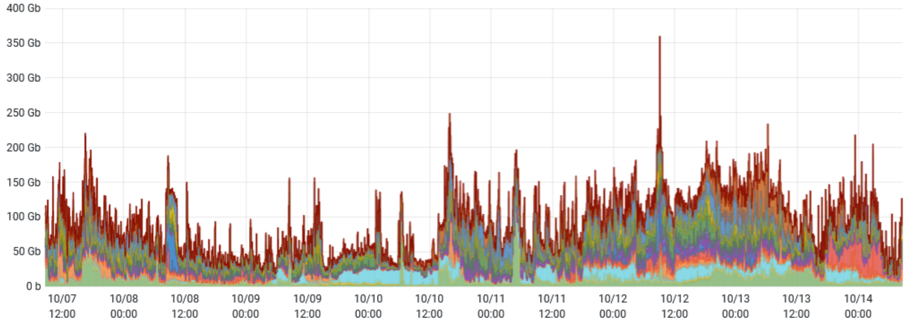
LHCOPN – traffic levels for T1T1 data transfer

LHCOPN Total Traffic (CERN -> T1s)



	max	avg	current	total
Outgoing DE-KIT	184 Gb	83.6 Gb	81.4 Gb	169 Tb
Outgoing KR-KISTI	8.83 Gb	170 Mb	1.36 Mb	342 Gb
Outgoing RU-T1	20.8 Gb	6.09 Gb	6.14 Gb	12.3 Tb
Outgoing FR-IN2P3	80.7 Gb	20.5 Gb	2.06 Mb	41.4 Tb
Outgoing NDGF	41.0 Gb	5.77 Gb	24.3 Gb	11.6 Tb
Outgoing NL-T1	94.0 Gb	6.97 Gb	445 Mb	14.0 Tb
Outgoing TW-ASGC	9.62 Gb	1.77 Gb	1.04 Gb	3.56 Tb
Outgoing IT-INFN-CNAF	98.9 Gb	12.3 Gb	98.9 Gb	24.8 Tb
Outgoing UK-RAL	22.9 Gb	6.13 Gb	18.4 Mb	12.4 Tb
Outgoing CA-TRIUMF	56.5 Gb	6.37 Gb	17.6 Gb	12.8 Tb
Outgoing US-BNL	95.9 Gb	15.7 Gb	37.9 Gb	31.6 Tb
Outgoing US-FNAL	56.9 Gb	5.87 Gb	1.12 Mb	11.8 Tb
Outgoing ES-PIC	12.1 Gb	3.13 Gb	0 b	6.31 Tb

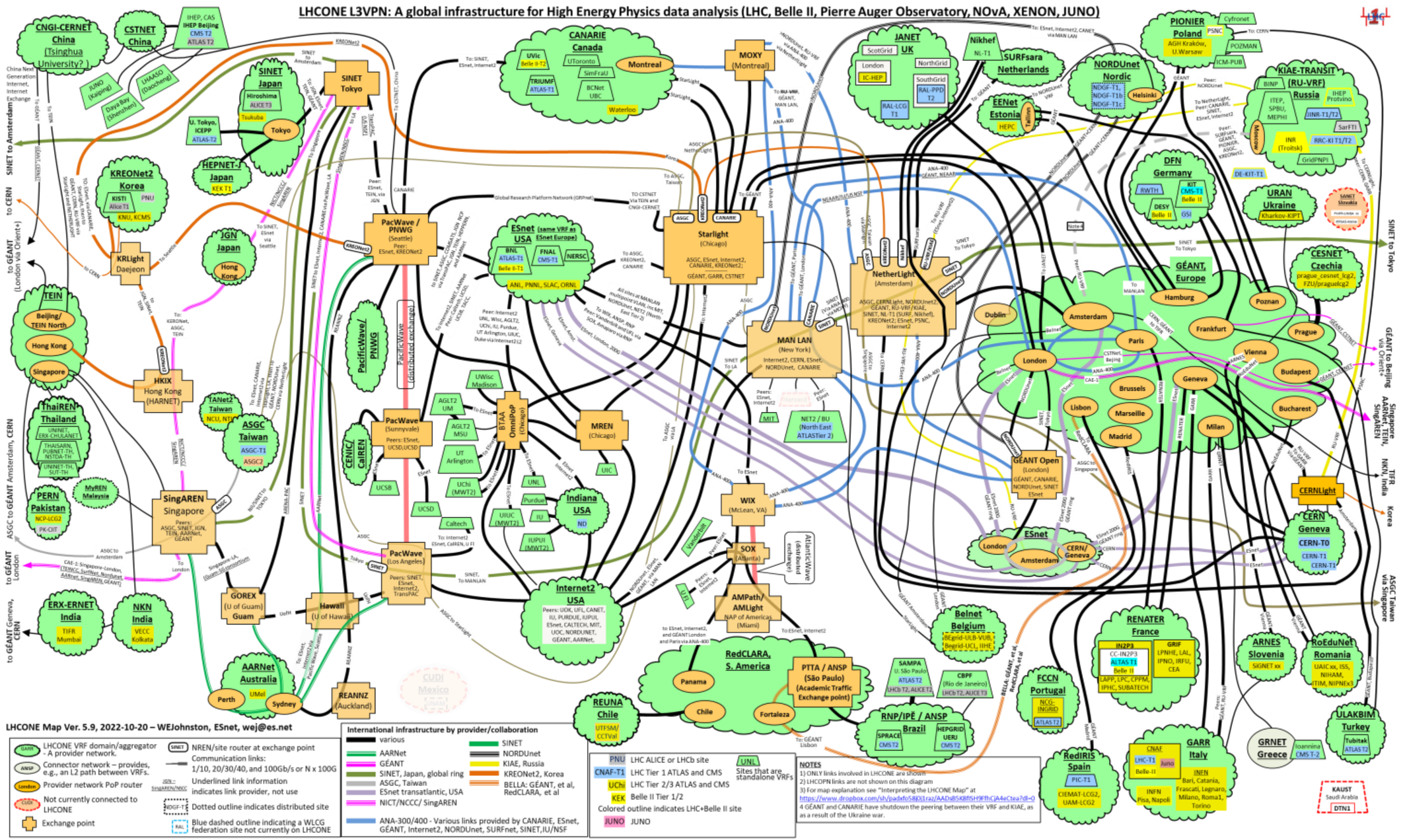
LHCOPN Total Traffic (T1s -> CERN)



	max	avg	current	total
Incoming DE-KIT	145 Gb	10.9 Gb	4.16 Gb	22.0 Tb
Incoming KR-KISTI	3.07 Gb	428 Mb	10.8 kb	864 Gb
Incoming RU-T1	40.7 Gb	9.61 Gb	3.60 Gb	19.4 Tb
Incoming FR-IN2P3	84.2 Gb	6.33 Gb	34.1 kb	12.8 Tb
Incoming NDGF	79.4 Gb	7.98 Gb	16.0 Gb	16.1 Tb
Incoming NL-T1	96.9 Gb	5.43 Gb	10.3 Gb	10.9 Tb
Incoming TW-ASGC	130 b	76.2 b	69.7 b	154 kb
Incoming IT-INFN-CNAF	62.7 Gb	4.94 Gb	1.64 Gb	9.95 Tb
Incoming UK-RAL	43.0 Gb	11.5 Gb	966 Mb	23.3 Tb
Incoming CA-TRIUMF	52.8 Gb	5.78 Gb	38.4 Gb	11.6 Tb
Incoming US-NBL	83.0 Gb	17.1 Gb	11.8 Gb	34.5 Tb
Incoming US-FNAL	61.0 Gb	8.18 Gb	799 b	16.5 Tb
Incoming ES-PIC	11.0 Gb	3.45 Gb	0 b	6.95 Tb

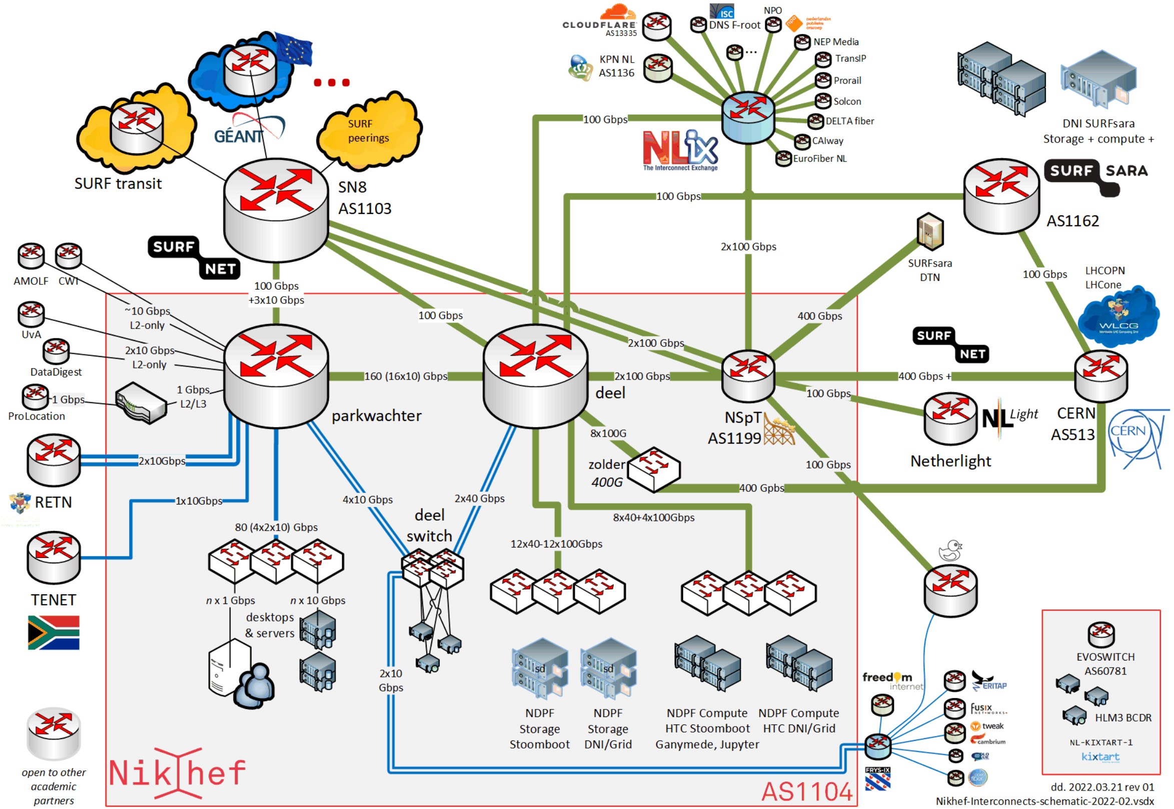
CERN OpenMonIT LHCOPN, period Oct 7 .. Oct 14 2022, from https://monit-grafana-open.cern.ch/d/HreVOyc7z/all-lhcopn-traffic

LHCone



LHCone ("LHC Open Network Environment") – visualization by Bill Johnston, ESnet version: October 2022 – updated with new AS1104 links

Just one random (smallish) autonomous system



AS1104

Exercising the network – sensor data and events

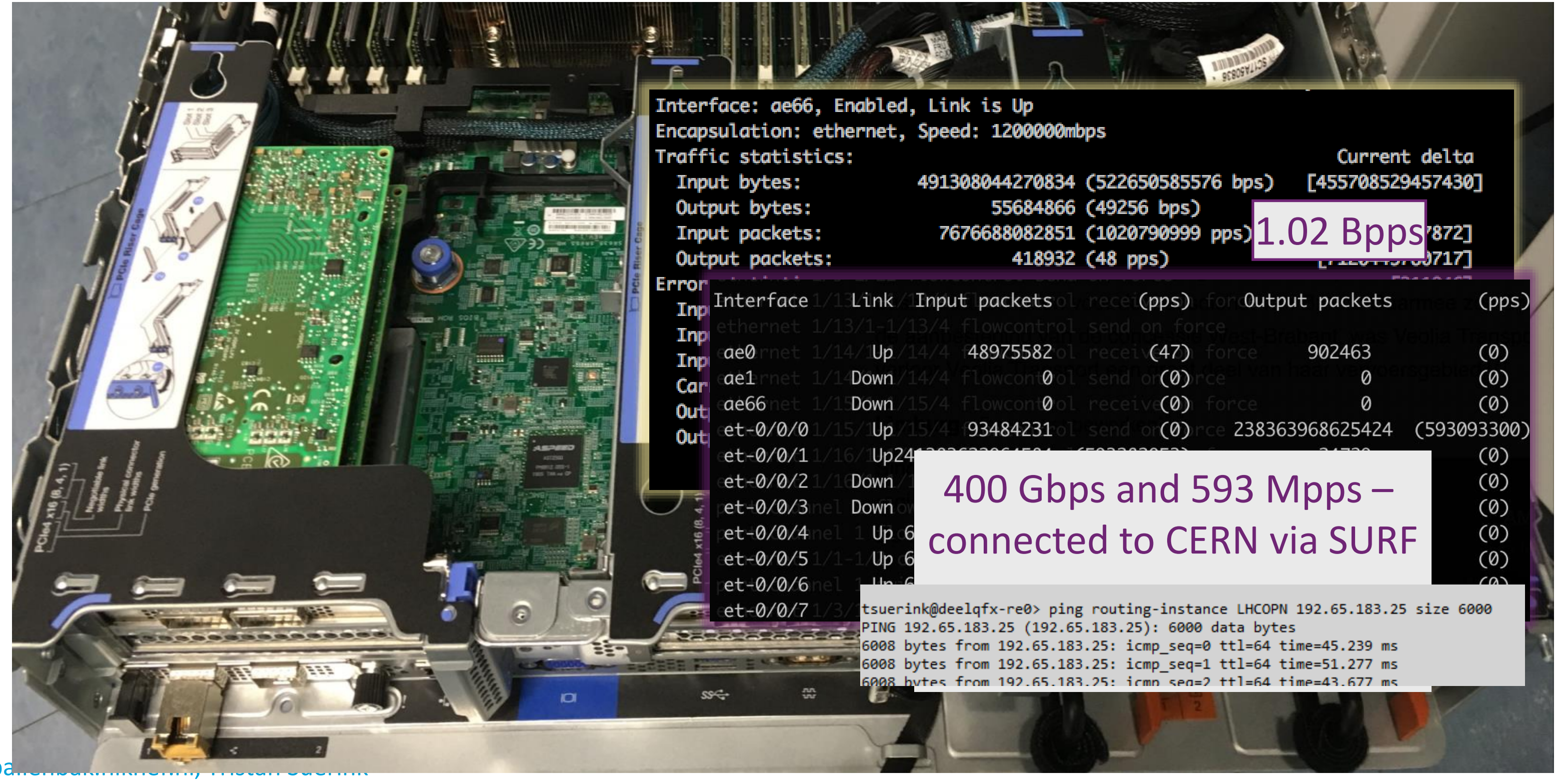
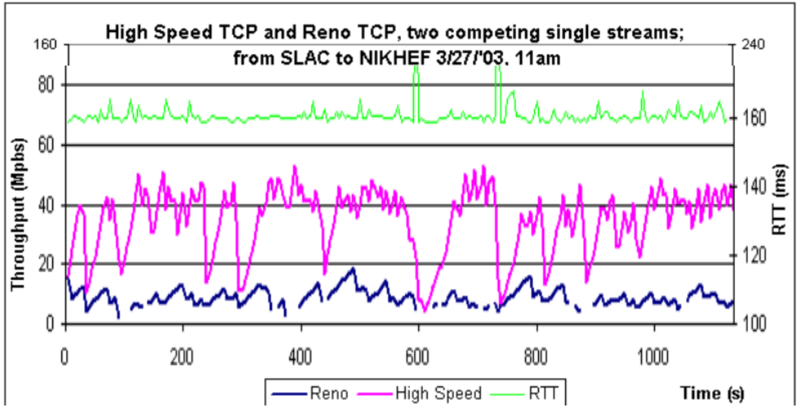


Image: ba

Scaling data access: 'system-aware design' at application layer

Reading data 'scattered' in a file - simply using POSIX-like IO - when done over the network severely exposes latency

and TCP slow-start makes that even worse



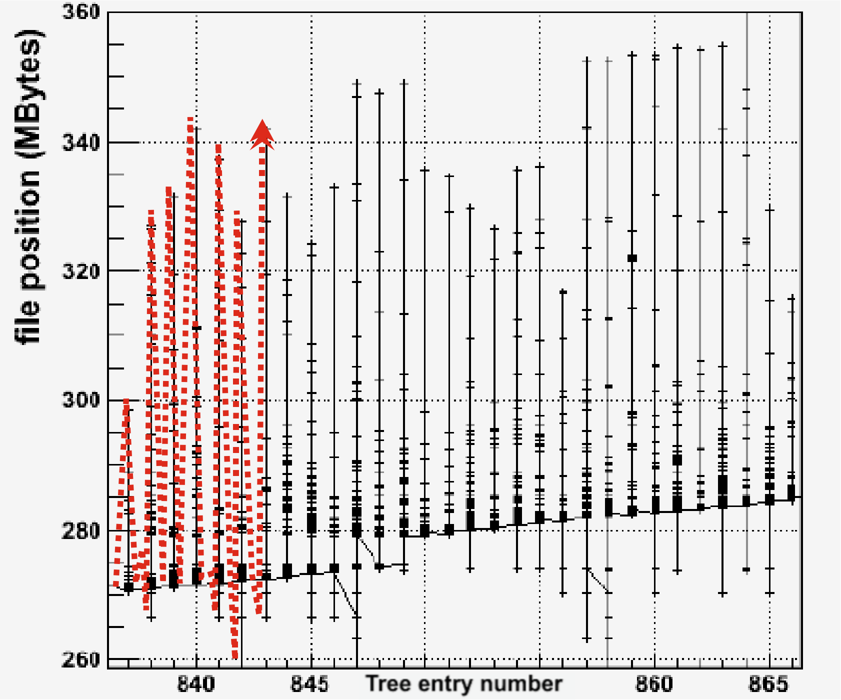


Image of TCP slow-start and packet loss impact (in Mpps): Antony Antony et al., Nikhef, for DataTAG, 2003(!)
Right: base graphic: Philippe Canal "Root I/O: the fast and the furious", CHEP2010 Access pattern reflects Root versions < 5.28, before Ttree caching and 'baskets'

Access, Trust & Identity

More than one user, *from*
more than one organizational domain, *in*
more than one country

WLCG: when we met a global trust scaling issue

- 170 sites
- ~60 countries & regions
- ~20000 users

just *how* many interactions

people photo: a small part of the CMS collaboration in 2017, Credit: CMS-PHO-PUBLIC-2017-004-3; site map: WLCG sites from Maarten Litmaath (CERN) 2021

Access control in a single domain

- Dedicated to each service where you need access
- Usually strongly linked to authorization: at times even different accounts for different roles
- In a multi-organizational system becomes

$$\mathcal{O}(n_{sites} * n_{services}) * \mathcal{O}(n_{users})$$

Image: AARC NA2 training module "Authentication and Authorisation 101" - https://aarc-community.org/training/aai-101/

Scaling issues – credentials at each site does not work

state of EDG and the HEP LHC computing in 2000

NIKHEF NATIONAAL INSTITUUT VOOR KERNFYSICA EN HOGE-ENERGIEFYSICA

Guest / students form (pleas

1. This form is completed in connection with: ☐ work experien ☐ otherwise, vi

CERN/User Registration

CERN COMPUTER CENTRE - US

http://cern.ch/it/documents/ComputerUsage/CompA

To be returned to the User Registration box at the en
completed by a user who requires a computer accou
Department, and is not yet registered in another gro

To be completed by the User :

It is MANDATORY to provide the following infor
treated confidentially and only be used for ensuring
Supply name as registered by the Users' Off
FAMILY NAME(S): ……………………………
FIRST NAME(S) : ……………………………………………………………………
SEX [M] [F] BIRTHDATE: Day …..… Month ……… Year ……………..
HOME INSTITUTE/FIRM: ……………………………………………………………….
NATIONALITY: ………………...*CERN SUPERVISOR…………………………….
*CERN DEPARTMENT: . . . . . *CERN ID NUMBER (as on CERN card)…………..

To be completed by the Group Administrator:

Fermilab

For Office Use Only

ID:	Action:	ID Exp:	
Insurance:	Medical:	Safety:	
Computer:	Stkrm:	Family:	
NON-473:	Sensitive:	Verifier:	Date:

Name:

Last	First	Middle
SWIETZER	JOHN	JAMES

University or Institution Name:	Telephone:
FLORIDA STATE UNIVERSITY	850-644-XXXX

Experiment/Department:

Exp. / Dept.	Spokesperson	Home Institution Contact	Contact Telephone
D0	WOMERSLEY/WEERTS	SHARON HAGOPIAN	850-644-4777

Authentication – who are you

Authenticating to a single service is relatively simple

- per-service identity (username) and secrets (e.g. password or TOTP token)
- server-side: list of valid users and (hashed and hopefully salted) secrets

```
[root@kwark ~]# cat /etc/passwd
root:x:0:0:root:/root:/bin/bash
bin:x:1:1:bin:/bin:/sbin/nologin
daemon:x:2:2:daemon:/sbin:/sbin/nologin
adm:x:3:4:adm:/var/adm:/sbin/nologin
lp:x:4:7:lp:/var/spool/lpd:/sbin/nologin
sync:x:5:0:sync:/sbin:/bin/sync
shutdown:x:6:0:shutdown:/sbin:/sbin/shutdown
halt:x:7:0:halt:/sbin:/sbin/halt
```

```
root:$6$s8ciAG5gLuv2bPQS$6EcskgtKvQ.rHbif
davidg:$6$nDYcIez2Uaufbtlg$R1hS/Qjn0qYQZk
marianne:$6$p3CeevG6jfNDqZj1$HKHqUTnt2fEqQfkA/m5J3oAOA0zSvgLCKOSQhPS
```

Passport image: cropped from original by Jon Tyson on Unsplash https://unsplash.com/photos/Hid-yhommOg

Authorization – what you are allowed to do

soon needs specifying **access rights** to resources, based on an access **policy**

- might be implicit or ad-hoc
- be in formal policy language
  example: Argus PDP
- or be service-specific
  example: Linux sssd config

```
resource "http://cern.ch/authz/ce1" {
   action "http://cern.ch/authz/actions/ce-submit" {
      rule permit {
         vo="atlas"
         pilot-job="true"
      }
      rule deny {
         pilot-job="true"
      }
   }
}
```

```
ldap_access_order = filter,authorized_service
ldap_access_filter = (|(memberOf=cn=gridSrvAdministrators,ou=DirectoryGroups,dc=farmnet,
dc=nikhef,dc=nl)(memberOf=cn=gridMWSecurityGroup,ou=DirectoryGroups,dc=farmnet,dc=nikhef
,dc=nl)(memberOf=cn=nDPFPrivilegedUsers,ou=DirectoryGroups,dc=farmnet,dc=nikhef,dc=nl))
```

Policy example: Argus system, https://argus-documentation.readthedocs.io/en/stable/misc/examples.html; service-specific: sssd.conf ldap auth_provider

Assertions to meet an authorization policy

assertions can be added to identity info

- e.g. visa are strongly bound
  to a specific entity
  through an identity statement by a trusted third party

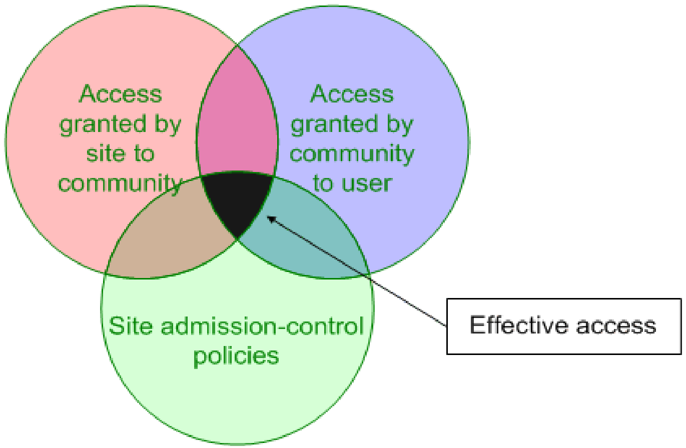
- but some are just a 'bearer token'

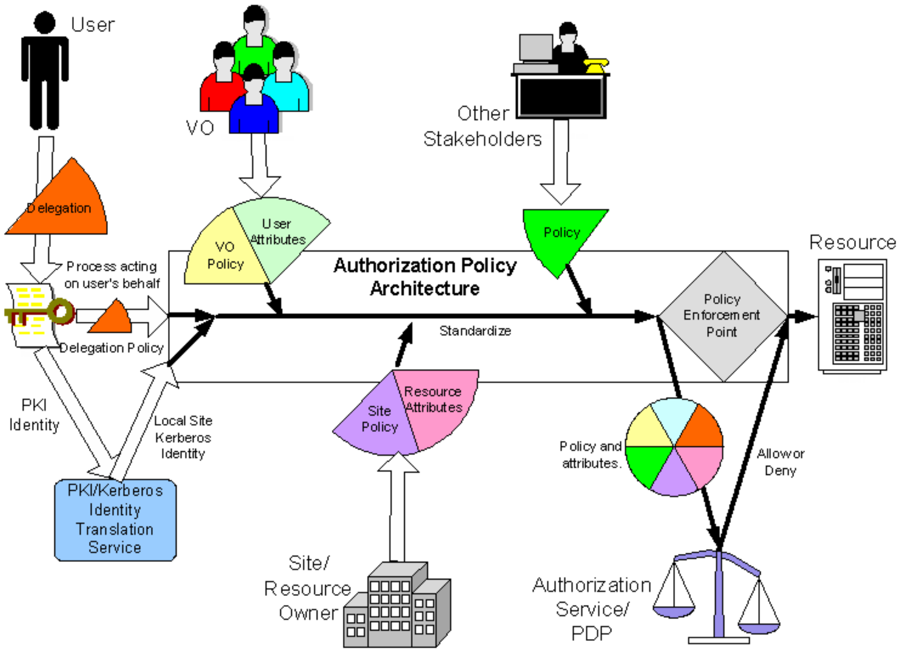
- others are looked up as needed

USA visa image source: https://2009-2017.state.gov/m/ds/rls/rpt/79785.htm

Authorization and access control

Unless data-level encryption is used, access control is ultimately enforced by the service provider





policy overlap diagram by Olle Mulmo, KTH for EGEE-I JRA3, policy pie: OpenGrod Forum OGSA working group and Globus Alliance

Authentication and Authorization Infrastructure

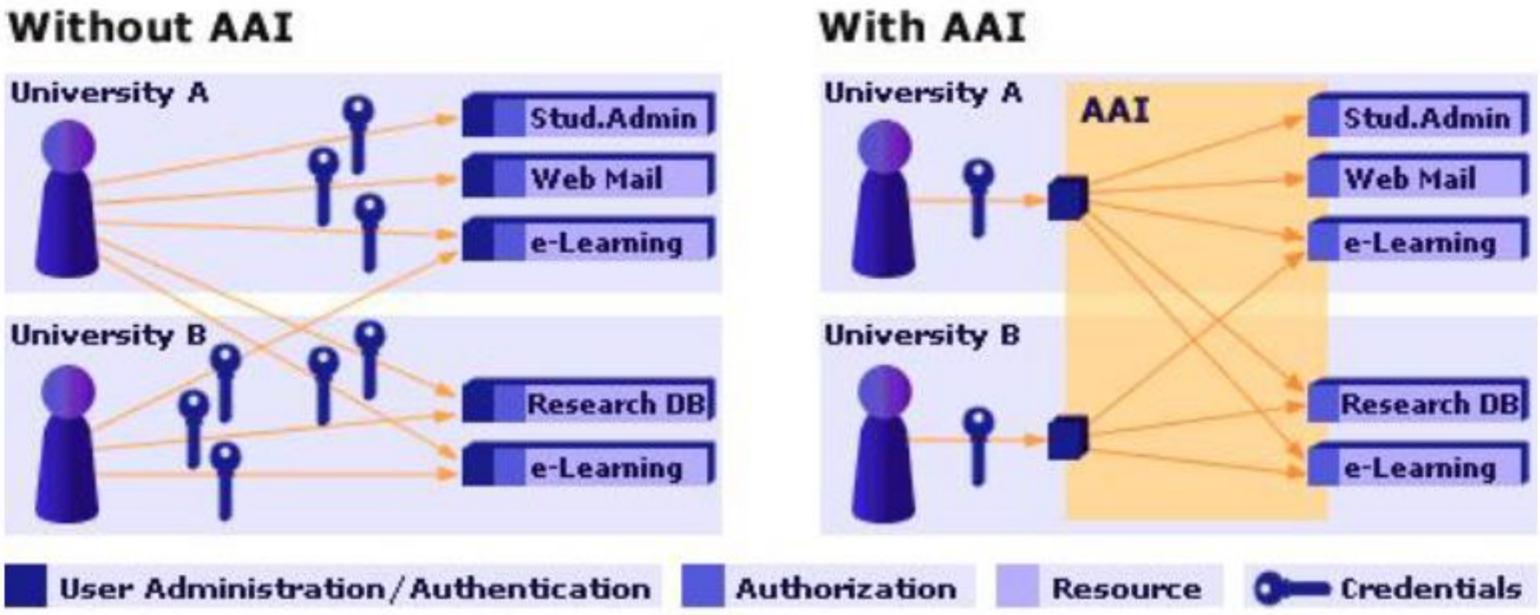
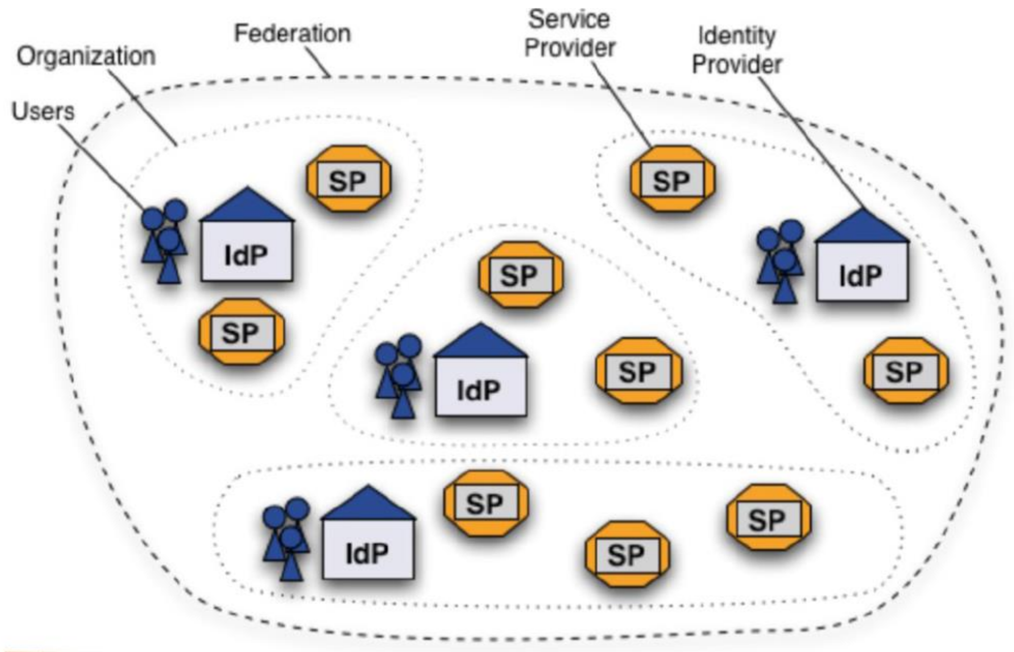


Image: AARC NA2 training module "Authentication and Authorisation 101" - https://aarc-community.org/training/aai-101/

Federation

portability of identity information across otherwise autonomous administrative domains

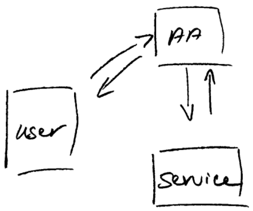


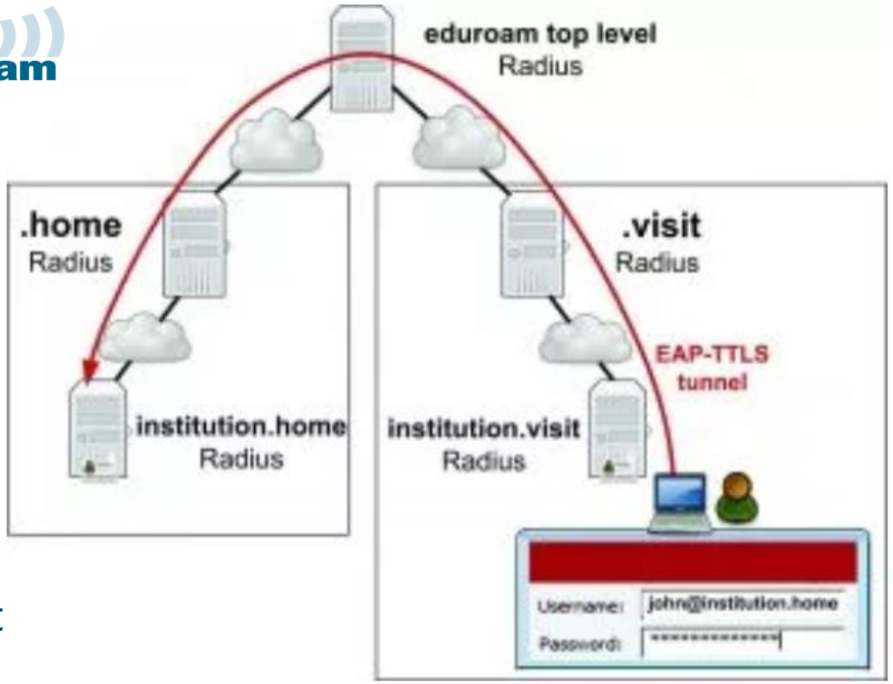
Shibboleth IdP image and SAML2 auth flow by SWITCH (CH) – see also https://refeds.org/ on federation structure and (assurance and security) guidelines

One simple federation you know: eduroam

service-specific trust between organisations globally

hierarchical RADIUS servers based an 802.1x secure exchange over TLS or EAP-TTLS tunneling your credentials back to your home institution





RADIUS server then instructs WiFi access point

eduroam: Klaas Wieringa et al., image from https://eduroam.org/how/, GEANT ; RADIUS: RC2865 https://www.rfc-editor.org/rfc/rfc2865; see also freeradius.org

skip past RSA background to 88

CIA interlude –trusted handshaking at a distance

Trust needs **C**onfidentiality, **I**ntegrity, and **A**vailability … and cryptography in some way

Client authentication

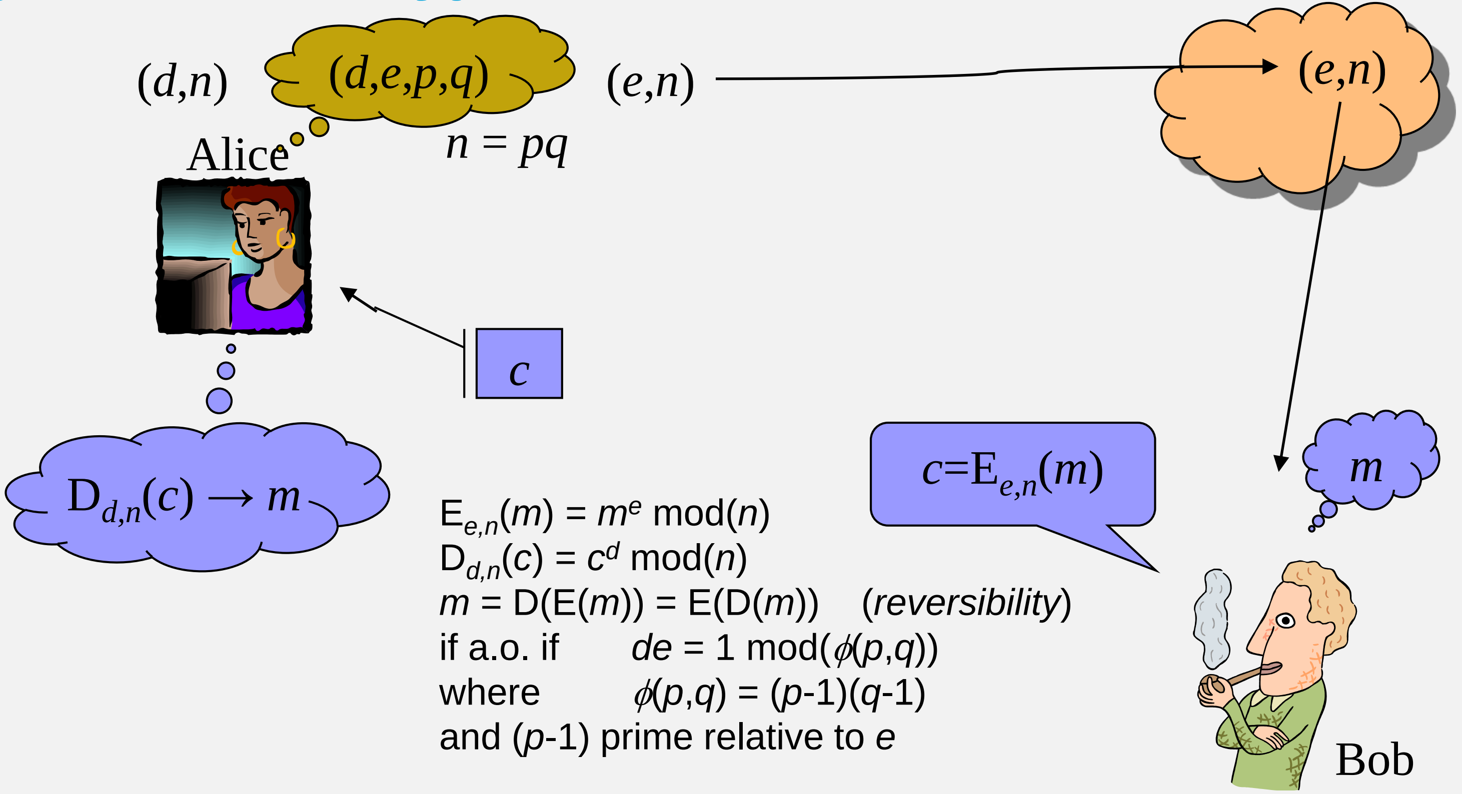
- pre-shared secrets, may be salted hashed on service side
- required: secure one-way hash function
- need a protected channel

Mutual authentication

- you either need lots of shared keys, or a trusted third party (TTP)
- with the TTP and multiple services comes the need for encryption
- across administrative domains, *key distribution* is the larger challenge

The cryptography used can be either *symmetric* or *asymmetric*, ‘public key’

Asymmetric crypto: RSA interlude needed?



$E_{e,n}(m) = m^e \bmod(n)$
$D_{d,n}(c) = c^d \bmod(n)$
$m = D(E(m)) = E(D(m))$ *(reversibility)*
if a.o. if $de = 1 \bmod(\phi(p,q))$
where $\phi(p,q) = (p\text{-}1)(q\text{-}1)$
and (p-1) prime relative to e

Rivest, Shamir and Adleman, Communications of the ACM 21 (2), 120-126

6-bit RSA (note: this might be broken quickly …)

- Take a (small) value e = **3**
- Generate a set of primes (p,q), each with a length of $k/2$ bits, with (p-1) prime relative to e.
  (p,q) = **(11,5)**
- $\phi(p,q)$ = (11-1)(5-1) = **40**; $n=pq=$**55**
- find d, in this case **27** [3*27 = 81 = 1 mod(40)]

- Public Key: **(3,55)**
- Private Key: **(27,55)**

$E_{e,n}(m) = m^e \bmod(n)$
$D_{d,n}(c) = c^d \bmod(n)$
$m = D(E(m)) = E(D(m))$ (*reversibility*)
if a.o. if $de = 1 \bmod(\phi(p,q))$
where $\phi(p,q) = (p-1)(q-1)$

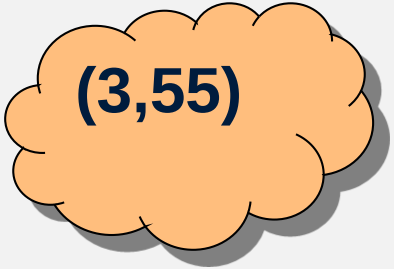
Message exchange

Encryption:

- Bob thinks of a plaintext $m(<n)$ = **18**
- Encrypt with Alice's public key **(3,55)**
- $c=E_{3;55}(18)=18^3 \text{ mod}(55) = 5832 \text{ mod}(55) =$ **2**
- send message **"2"**

Decryption:

- Alice gets **"2"**
- she knows private key **(27,55)**
- $E_{27;55}(2) = 2^{27} \text{ mod}(55) =$ **18** !



$E_{e,n}(m) = m^e \text{ mod}(n)$

$D_{d,n}(c) = c^d \text{ mod}(n)$

$m = D(E(m)) = E(D(m))$

if a.o. if $de = 1 \text{ mod}(\phi(p,q))$

where $\phi(p,q) = (p-1)(q-1)$

If you just have (3,55), it's hard to get the 27…

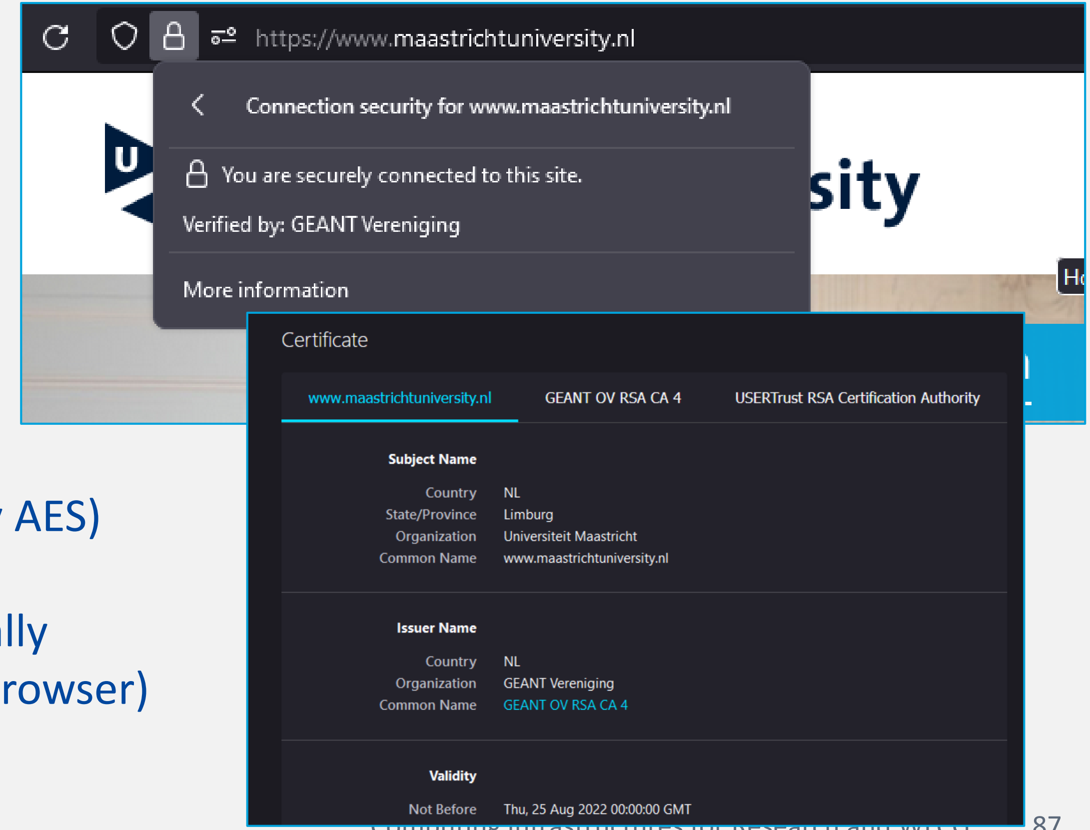
but also: the maximum plaintext is limited by the modulus length

The most used asymmetric crypto application

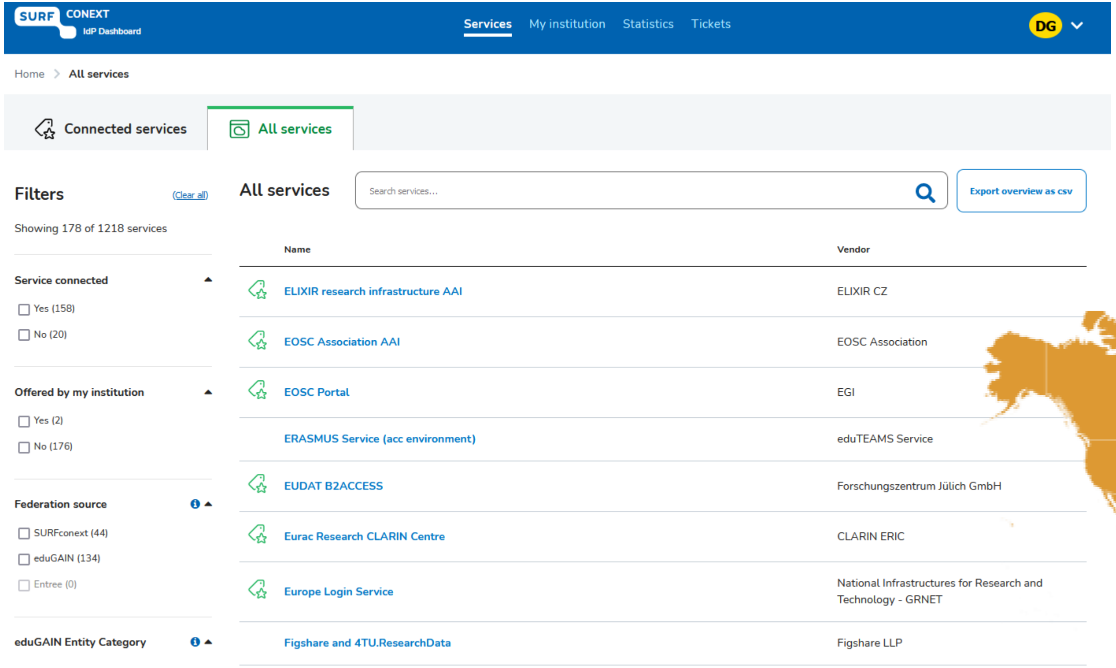
Asymmetric crypto underpins
the transport layer security
of all of the web today

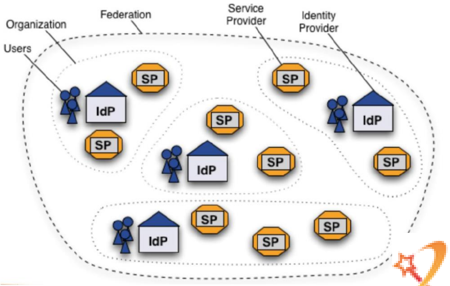
- ASN.1 syntax data with X.509 (RFC5280) structure
- mostly RSA or Elliptic Curves (EC)
- used to negotiate a (symmetric) bulk cipher (typically AES)

then used to protect channel to usually *unauthenticated* client application (browser)

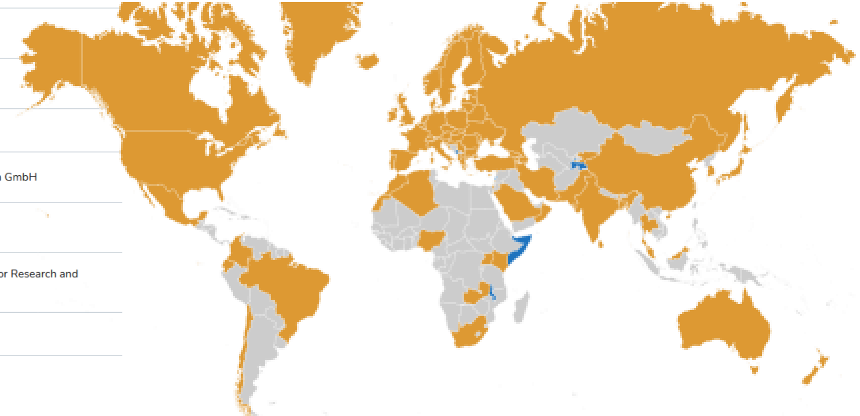


Multipurpose federation with SAML: SURFconext & eduGAIN



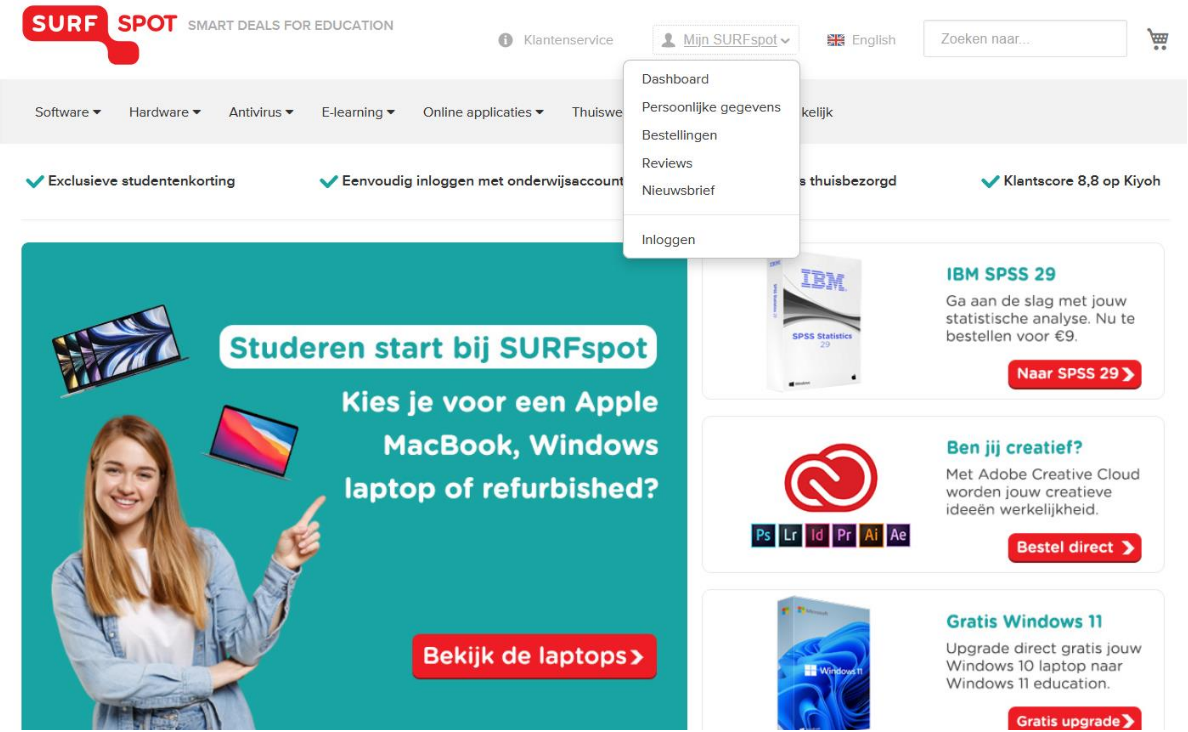


eduGAIN



Images: SURFconext IdP dashboard by SURF, showing some services tagged with REFEDS R&S; eduGAIN map: GEANT, https://technical.edugain.org/status

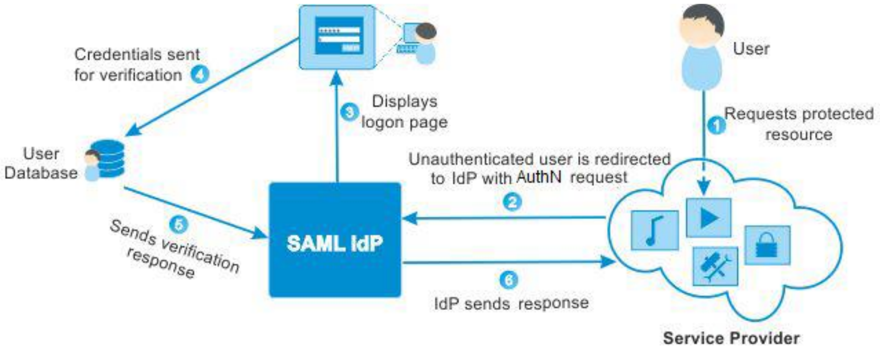
Your favourite federated service?



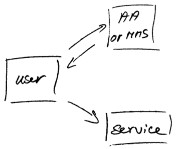
https://surfspot.nl/

SAML federation

Attributes	Values
E-mail	davidg@nikhef.nl
Affiliation	• employee • member • faculty
Targeted ID	https://sso.nikhef.nl/sso/saml2/idp/metadata.php!https://attribute-viewer.aai.switch.ch/shibboleth!b9f858169ea28dc68b6753baa1084d8c039e36a7
Common Name	David Groep
Display Name	David Groep
Principal Name	davidg@nikhef.nl
Home organization (international)	nikhef.nl
Home organization type (international)	urn:mace:terena.org:schac:homeOrganizationType:int:other



SAML2.0 auth flow



Try at https://attribute-viewer.nikhef.nl/ and select “Login via a global authentication SAML source”

Firefox: use F12, and SAML message decoder: https://addons.mozilla.org/en-US/firefox/addon/saml-message-decoder-extension/ (Magnus Suther)

SAML WebSSO flow image: SWITCH, CH

Under the hood, it’s a (signed) XML document

```
<saml:Subject>
     <saml:NameID Format="urn:oasis:names:tc:SAML:2.0:nameid-format:persistent">xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx</saml:NameID>
     <saml:SubjectConfirmation Method="urn:oasis:names:tc:SAML:2.0:cm:bearer">
       <saml:SubjectConfirmationData NotOnOrAfter="2022-10-21T18:16:40Z"
         Recipient="https://attribute-viewer.aai.switch.ch/Shibboleth.sso/SAML2/POST"
         InResponseTo="_64c10a60c382bdaeb328653d9d25951c" /></saml:SubjectConfirmation>
   </saml:Subject>
   <saml:Conditions NotBefore="2022-10-21T18:11:39Z"
                    NotOnOrAfter="2022-10-21T18:16:40Z">
     <saml:AudienceRestriction>
       <saml:Audience>https://attribute-viewer.aai.switch.ch/shibboleth</saml:Audience>
     </saml:AudienceRestriction>
   </saml:Conditions>
   <saml:AuthnStatement AuthnInstant="2022-10-21T17:33:29
                        SessionNotOnOrAfter="2022-10-22T0
                        SessionIndex="_90f745f18f712b6a56
    <saml:AuthnContext>
        <saml:AuthnContextClassRef>urn:oasis:names:tc:SAM
        <saml:AuthenticatingAuthority>https://sso.nikhef.
    </saml:AuthnContext>
   </saml:AuthnStatement>
```

```
<saml:AttributeStatement>
      <saml:Attribute Name="urn:mace:dir:attribute-def:cn"
                      NameFormat="urn:oasis:names:tc:SAML:2.0:attrname-format:uri">
        <saml:AttributeValue xsi:type="xs:string">David Groep</saml:AttributeValue>
      </saml:Attribute>
      <saml:Attribute Name="urn:oid:2.5.4.3"
                      NameFormat="urn:oasis:names:tc:SAML:2.0:attrname-format:uri">
        <saml:AttributeValue xsi:type="xs:string">David Groep</saml:AttributeValue>
      </saml:Attribute>
      <saml:Attribute Name="urn:mace:dir:attribute-def:eduPersonAffiliation"
                      NameFormat="urn:oasis:names:tc:SAML:2.0:attrname-format:uri">
        <saml:AttributeValue xsi:type="xs:string">employee</saml:AttributeValue>
        <saml:AttributeValue xsi:type="xs:string">member</saml:AttributeValue>
        <saml:AttributeValue xsi:type="xs:string">faculty</saml:AttributeValue>
      </saml:Attribute>
      <saml:Attribute Name="urn:oid:1.3.6.1.4.1.5923.1.1.1.1"
      ...
```

Federation: different technologies, same idea

SAML - Security Assertion Markup Language and WebSSO (‘SAML2Int’)

- XML-formatted ‘attribute statements’ over web transport (usually POST)
- SAML-Metadata: list of entities with description of bindings with entityAttributes

PKI - Public Key Infrastructures

- certification authority (CA) signing X.509 formatted certificates
  with name, issuer, serial number, and extensions
- CAs can sign end-entities as well as other CAs (hierarchically or by cross-signing)
- bridge CAs render a technical implementation of a shared policy (assurance)
- policy-bridges don’t sign anything, but curate *distribution* (like browsers and operating systems based on CA/BF requirements, or the IGTF for research infras)

OIDC Fed - OpenID Connect Federation

- for end-points for OIDC Providers and Relying Parties – otherwise quite similar

federation based on ‘ultimate trust’ domains (e.g. cross-realm Kerberos) also exists, but …

See www.oasis.org for SAML, RFC5280 (tech) & RFC3247 (policy) for PKIX, https://igtf.net/ and https://cabforum.org;
OpenID Connect Federation: https://openid.net/specs/openid-connect-federation-1_0.html

Federation: technology, interoperability, policy

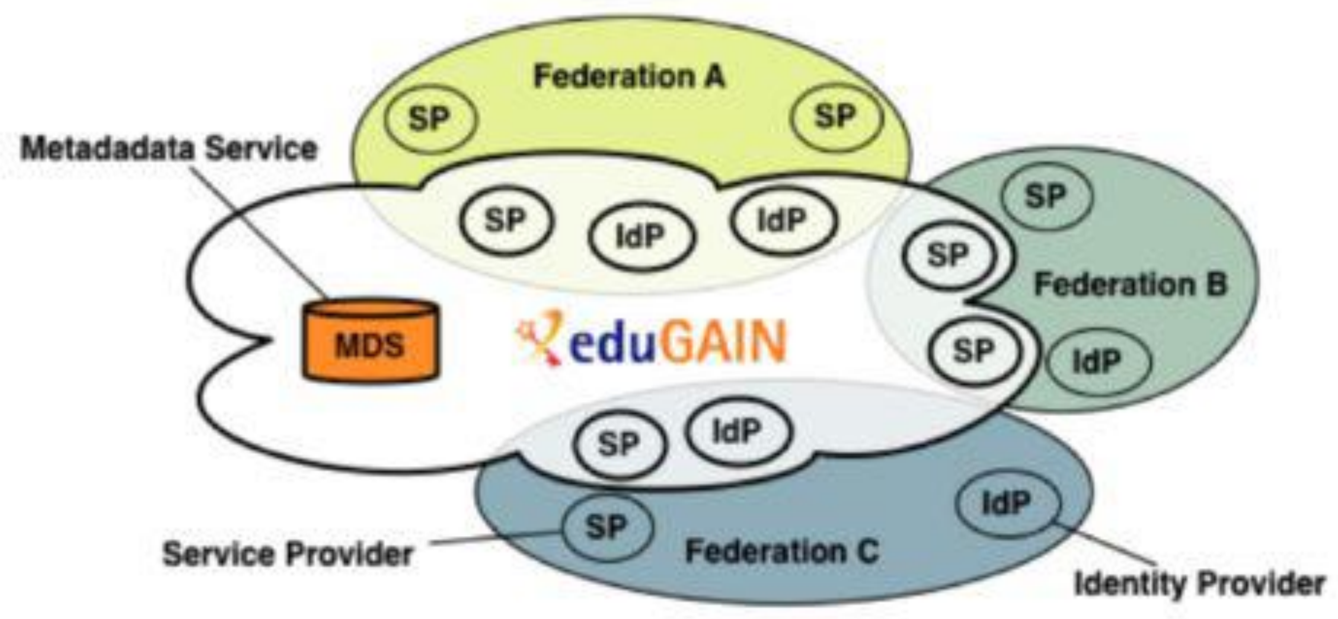
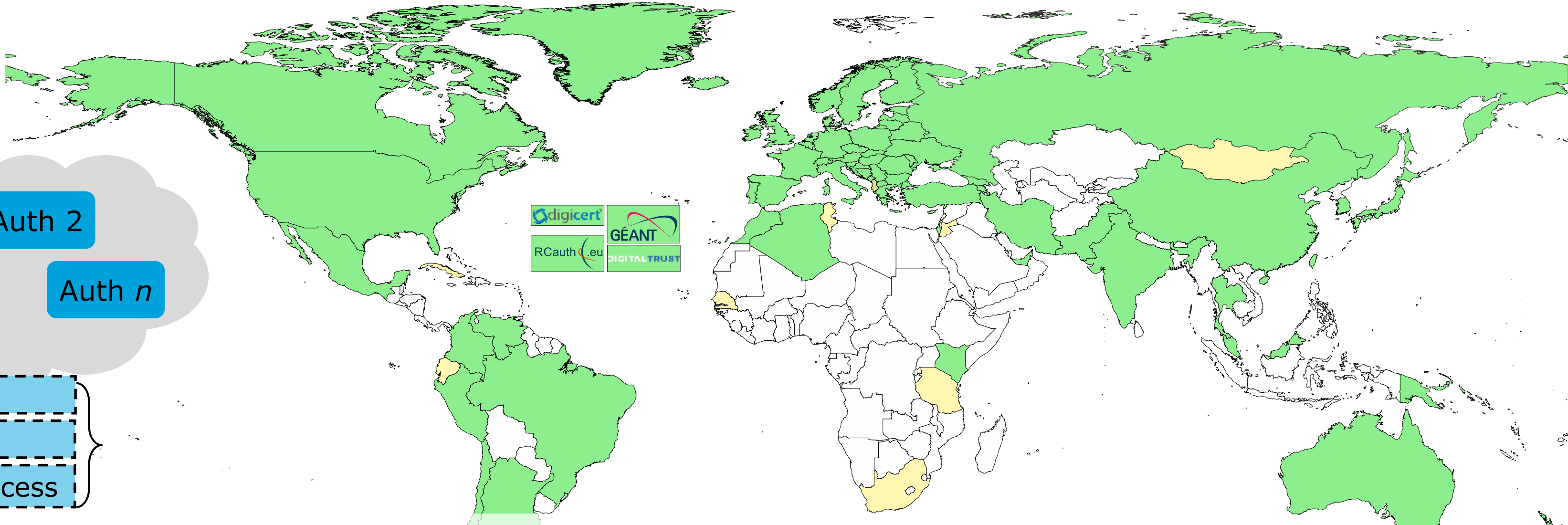


Image from SWITCH (CH) and edugain.org

Policy-bridged global federations for research computing



Authority 1
Auth 2
Auth *n*
Auth 3

charter
guidelines
acceptance process

relying party 1
relying party *n*

IGTF
AP | EU | TAG

3 regional IGTF chapters: EMEA, Americas, Asia Pacific

~ 90 Identity Providers (some leveraging a R&E federation)

~ 10 international major relying parties

~ 60 countries / economic areas / international treaty orgs

> 1000 relying service provider collaborations

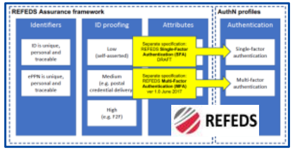
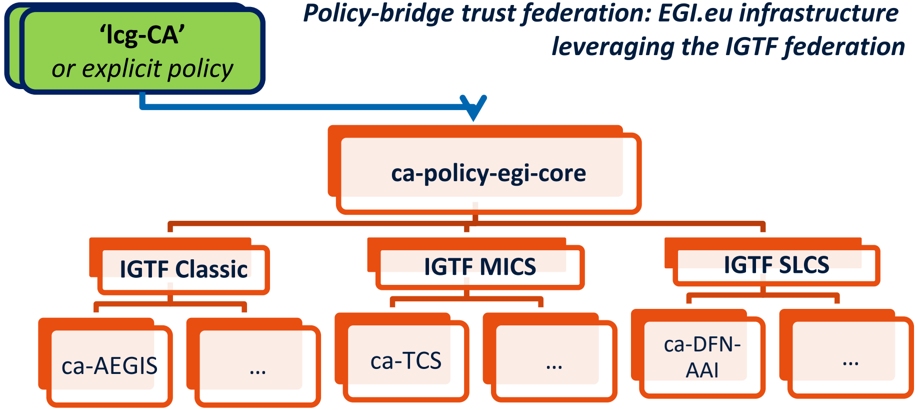


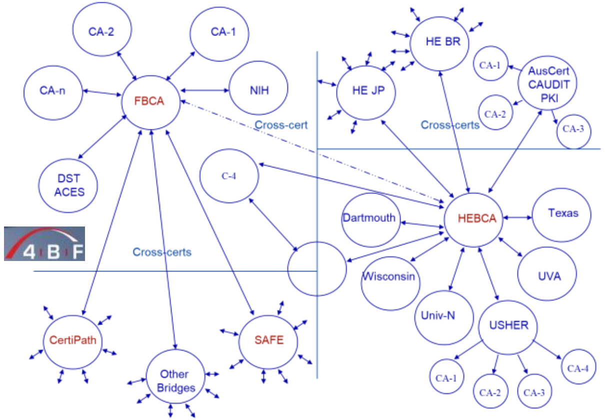
Image: Interoperable Global Trust Federation IGTF, https://igtf.net/; REFEDS Assurance Framework RAF: http://refeds.org/assurance, https://refeds.org/profile/mfa

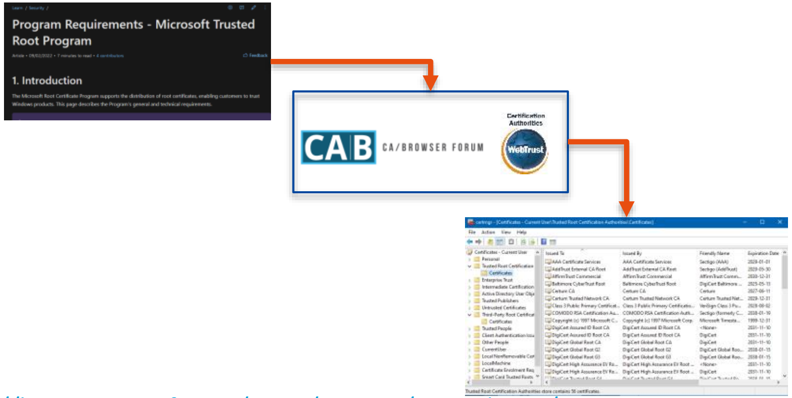
PKIX federation

trust remains with the relying party
can be *bridged* by either cross-signing (left) or by policy agreements (right)

Policy-bridge trust federation: EGI.eu infrastructure leveraging the IGTF federation





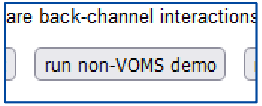


Left-hand image: 4 Bridges Forum, source: Scott Rea (then: Dartmouth)
Images: cabforum.org, WebTrust logo: from DigiCert.com; image MS root store, https://learn.microsoft.com/en-us/security/trusted-root/program-requirements

An X.509 RFC5280 Certificate (textually)

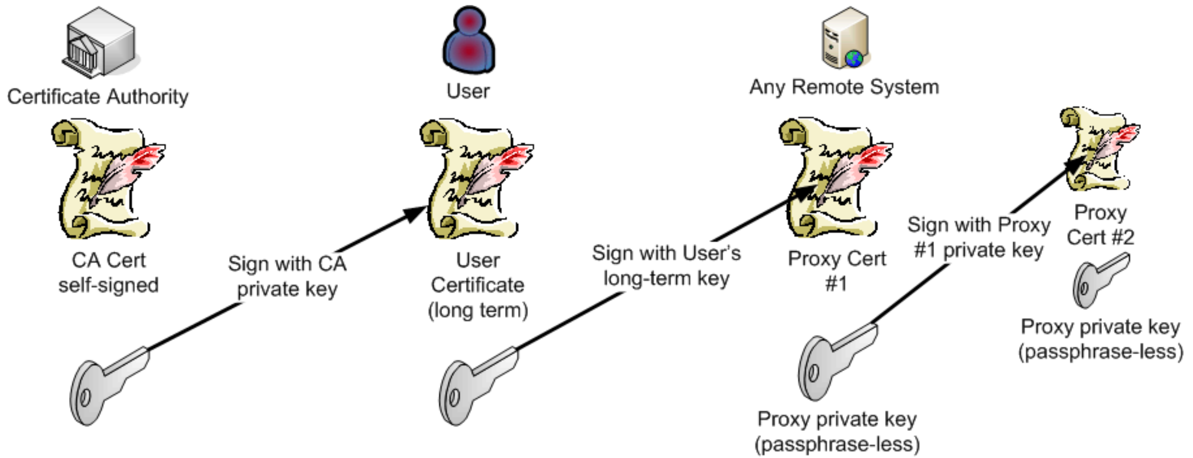
```
Version: 3 (0x2)
Serial Number:
    34:f3:e3:5f:c0:53:0b:a6:ef:2b:4a:79:01:b5:50:3b
Signature Algorithm: sha384WithRSAEncryption
Issuer: C = NL, O = GEANT Vereniging, CN = GEANT eScience Personal CA 4
Validity
    Not Before: Apr  2 00:00:00 2022 GMT
    Not After : May  2 23:59:59 2023 GMT
Subject: DC = org, DC = terena, DC = tcs, C = NL, O = Nikhef, CN = David Groep davidg@nikhef.nl
Subject Public Key Info:
    Public Key Algorithm: rsaEncryption
        RSA Public-Key: (4096 bit)
        Modulus:
            00:f0:0d:c0:ff:ee:f0:0d:f0:0d:c0:ff:ee:f0:0d:
            ...
            ff:50:6d
        Exponent: 65537 (0x10001)
X509v3 extensions:
    X509v3 Key Usage: critical
        Digital Signature, Key Encipherment
    X509v3 Basic Constraints: critical
        CA:FALSE
    X509v3 Extended Key Usage:
        E-mail Protection, TLS Web Client Authentication
    X509v3 Certificate Policies:
        Policy: 1.2.840.113612.5.2.2.5
```

You should be able to get a 'DOGWOOD' assurance certificate from RCauth.eu. Go to https://rcdemo.nikhef.nl/ and select the 'Basic demo' and use 'run non-VOMS' to get and view your short-lived certificate



PKIX certificates (and proxies for non-web access)

- Certificates are ASN.1 structures with (issuer, subject, serial) + extensions
- The digest (hash) signed with the private key of the issuer
- Verifiable using the issuer's public key

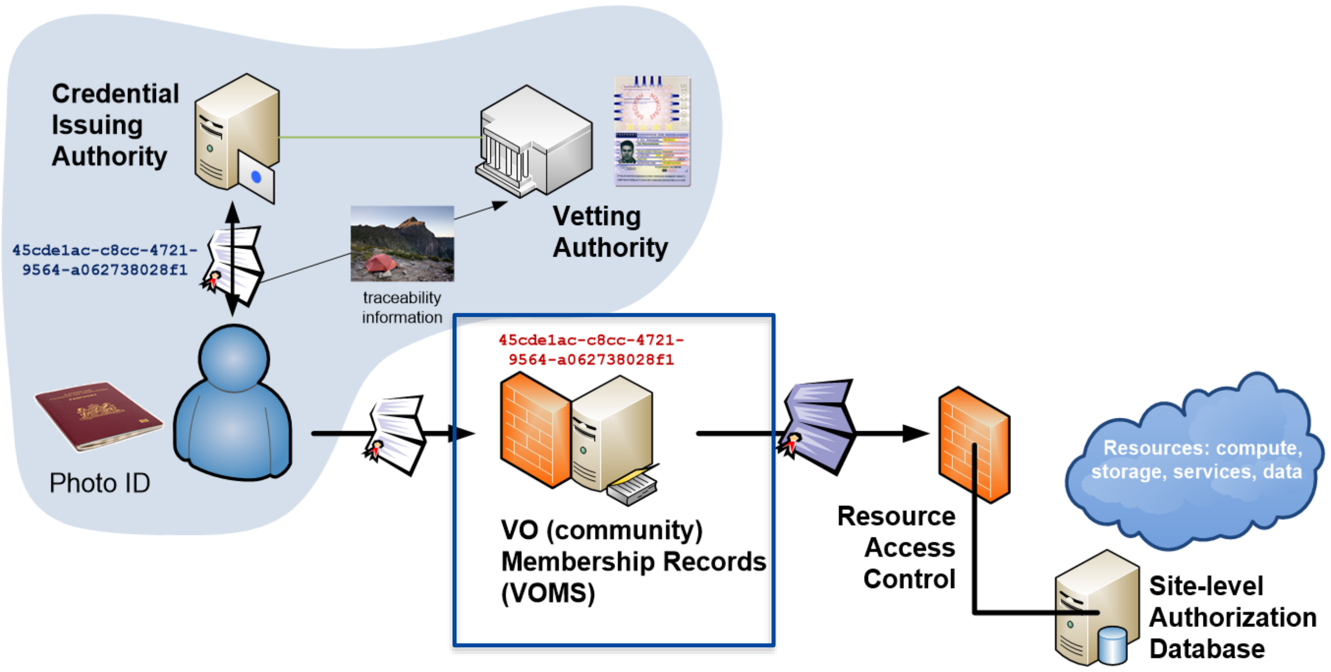


RFC3820 'proxy' certificates extend this concept to (restricted) identity delegation

To get an RFC3820 proxy certificate using your own federated identity, use RCauth.eu – see https://rcdemo.nikhef.nl/ and use the "Basic Demo" option

Identity federations give … identity (“AuthN”)

Authorization (what may you do) still needs to be added to the mix

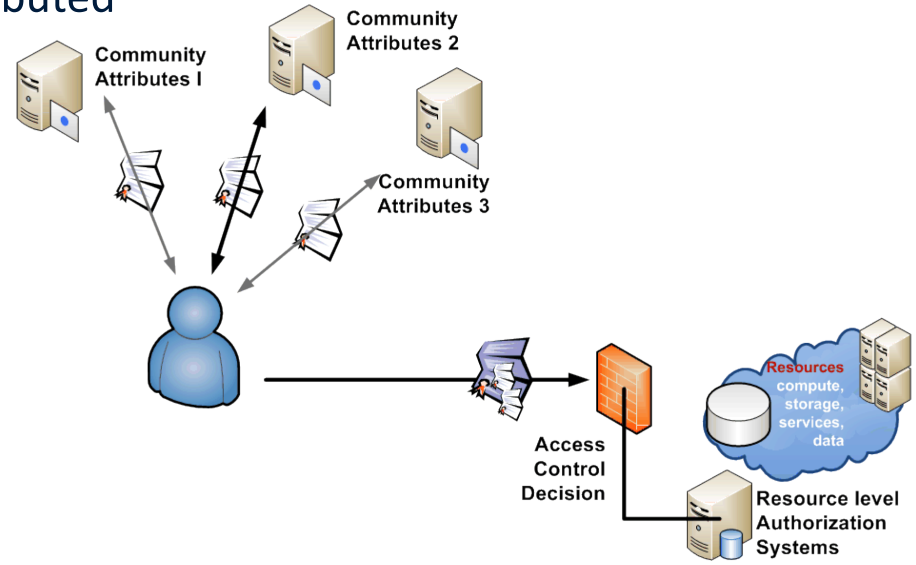


Multiple sources of authority: the community

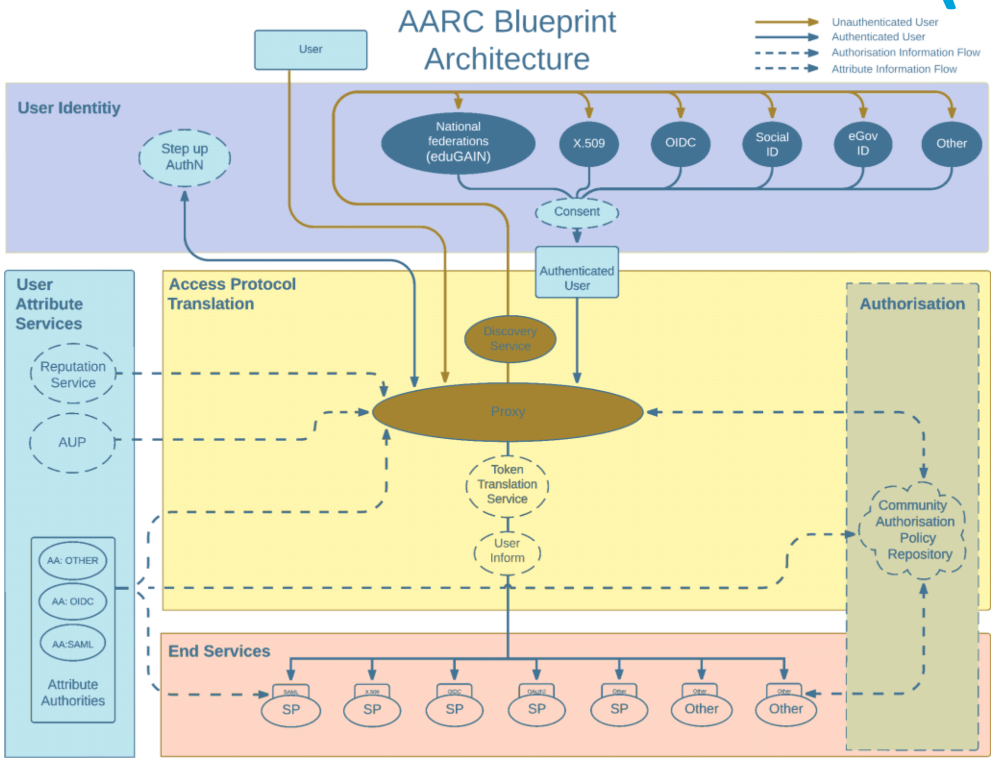
- authorization assertion providers (attribute authorities) use the identifier(s) from authentication in their membership services

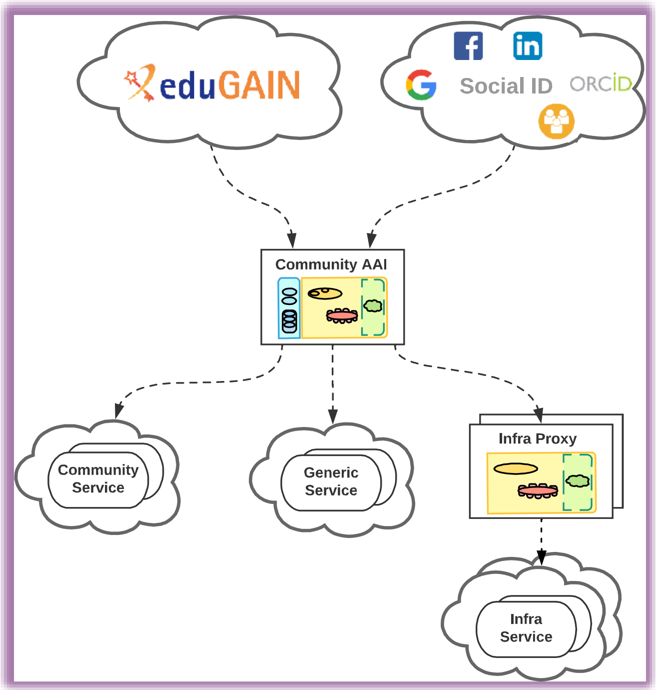
- *source of authority* for attributes is distributed

  e.g. community membership
  from the experiments,
  home affiliation from a university



Most trust flows from the (research) community





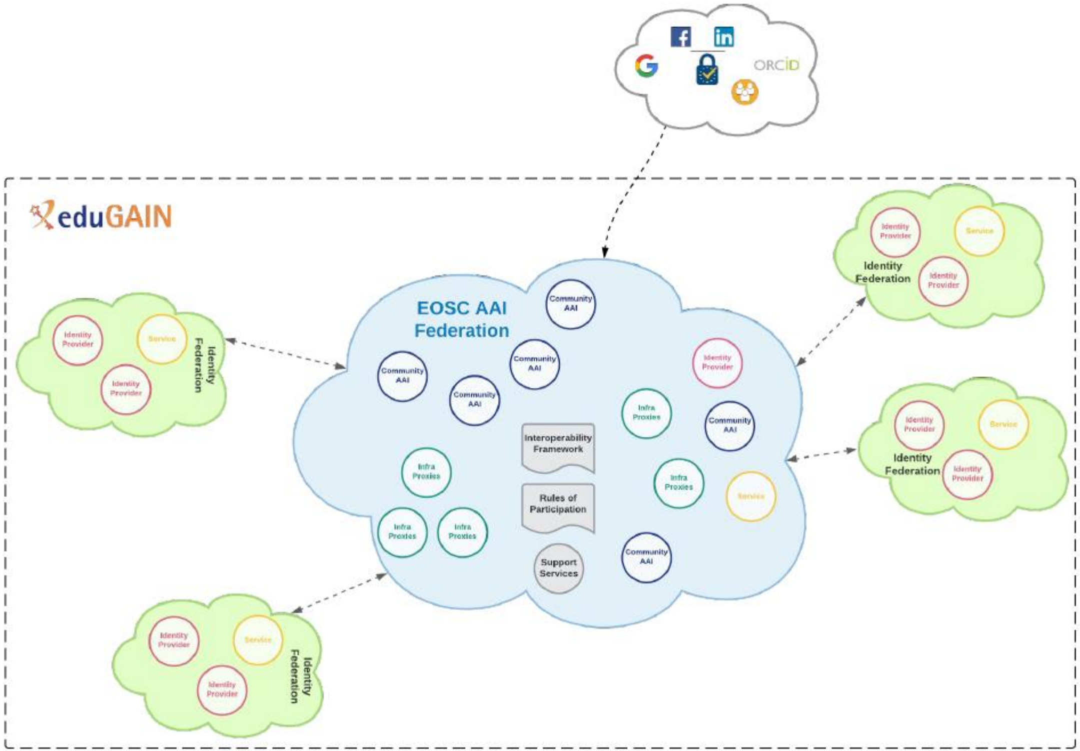
AARC Blueprint Architecture (2019) AARC-G045 https://aarc-community.org/guidelines/aarc-g045/; stacked proxies: EOSC AAI Architecture
EOSC Authentication and Authorization Infrastructure (AAI), ISBN 978-92-76-28113-9, http://doi.org/10.2777/8702

Example: European Open Science Cloud

eosc

EOSC Portal & Marketplace Amnesia service by the OpenAIRE e-infrastructure, EOSC Helpdesk: Zammad hosted by KIT https://eosc-helpdesk.eosc-portal.eu

EOSC AAI Federation – beyond the proxy again

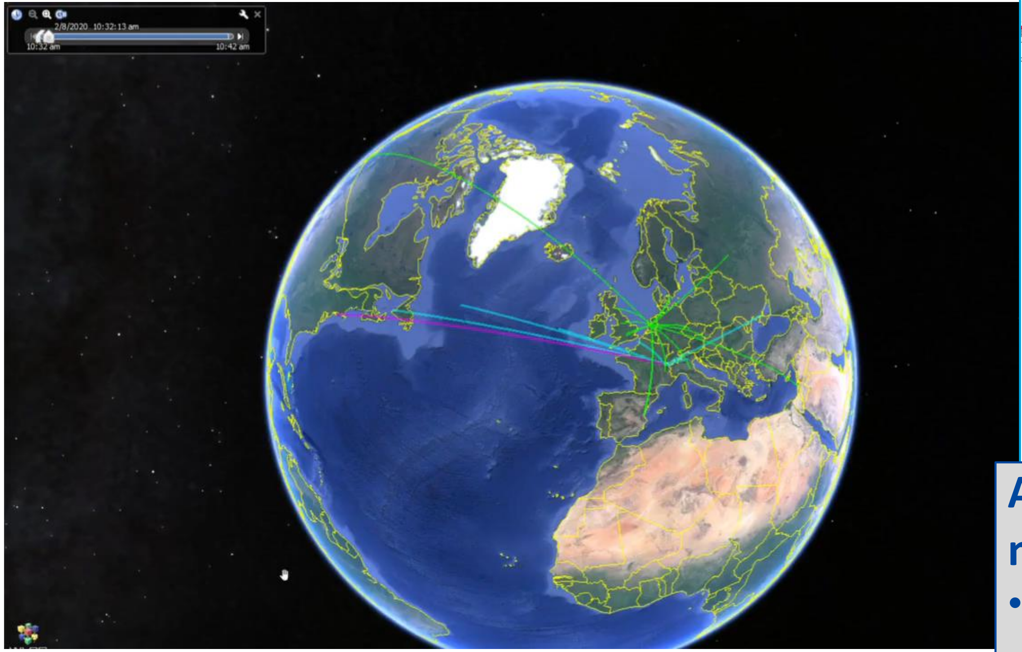


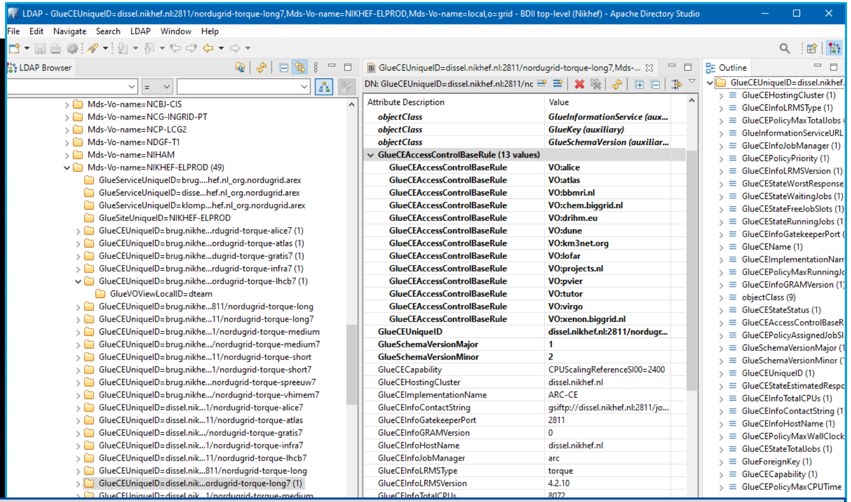
Christos Kanellopoulos (GEANT) for the EOSC AAI Federation in "The EOSC Core", https://eoscfuture.eu/wp-content/uploads/2022/04/EOSC-Core.pdf

Putting it back together again

Common patterns in scalability

A global infrastructure of EGI, OSG and WLCG, …





An infrastructure with components matched to application needs

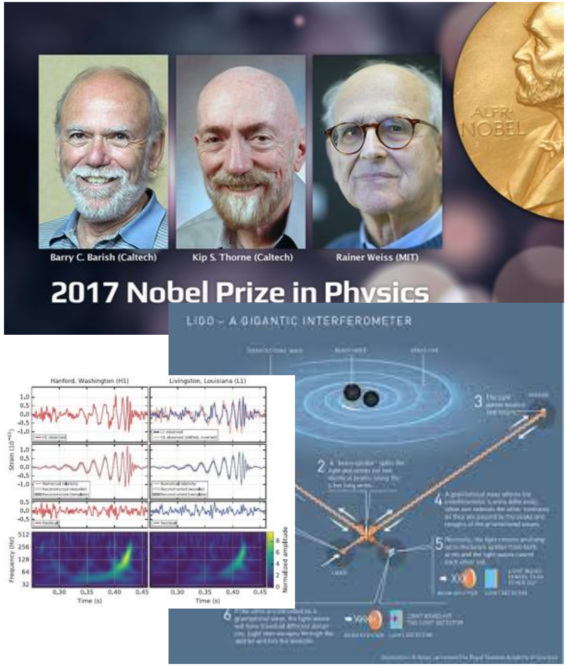
- systems architecture, compute (clusters), networking, storage, and application structure
- in a cost-efficient, and energy-efficient, way

BerkeleyDB Information System for EGI, from top-level BDII at ldap://bdii03.nikhef.nl:2170/o=grid; Earth visualization: https://dashb-earth.cern.ch/, Google Earth

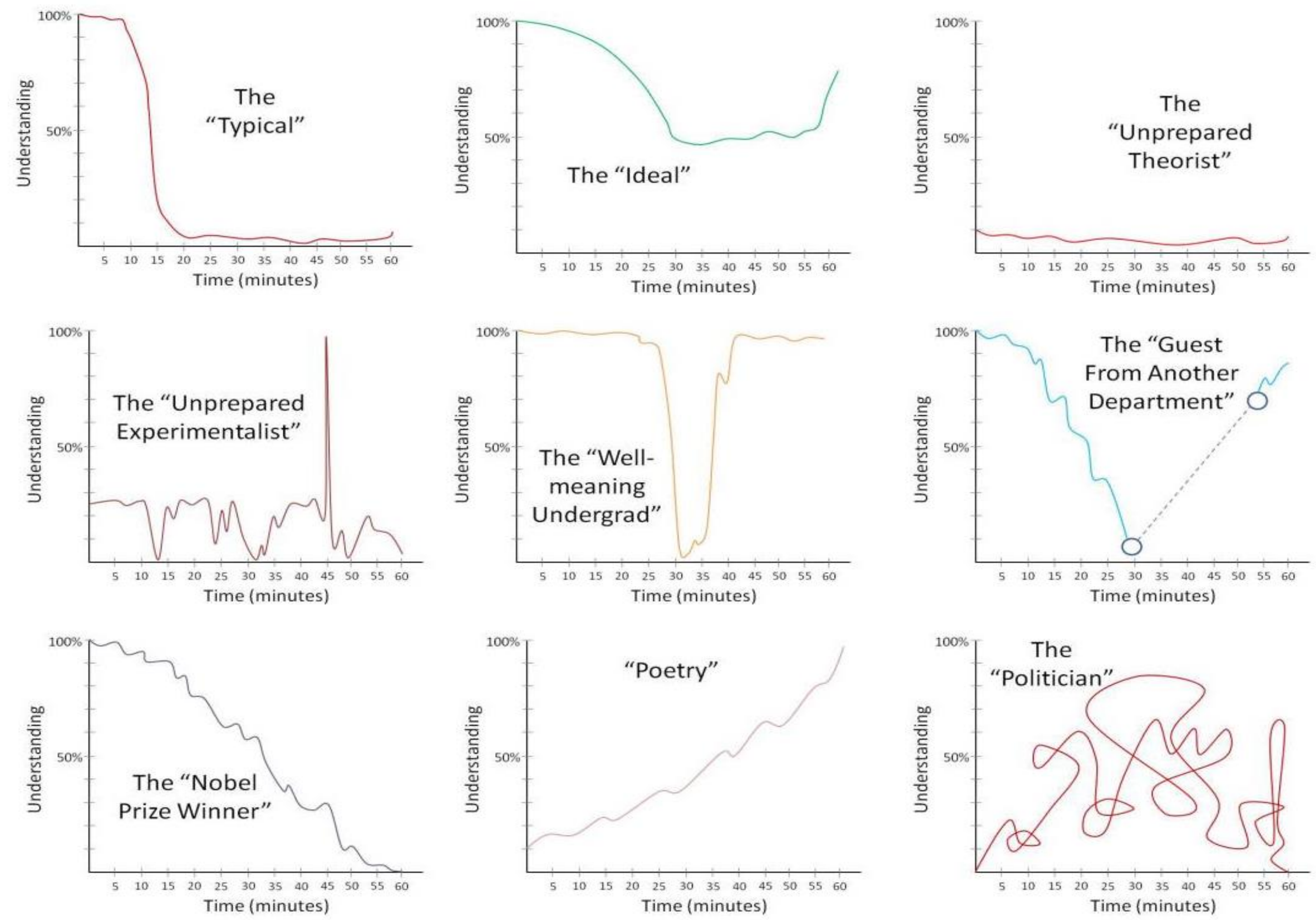
Did you discern a common pattern?

- Make central components passive and as stateless as possible
  - e.g. for fabric management, have central repository be a cacheable web service
  - although persistent storage obviously has to retain some state ☺
- Move complexity and volume requirements to the edge
  - the edge scales horizontally and scaling from 2+ is much easier than from 1→ 2
- You can move problems around, but it's hard to actually *solve* them
  - e.g. lack of a single common interface implies one needs adaptors and plugins
- Scaling *collaboration and trust* federation is as complex as scaling systems
  - and beyond 'Dunbar's Number', ~150, you will need some assessment and policy

e-Infrastructures & WLCG was one (of many) ingredients …



CERN Higgs discovery conference, with Fabiola Gianotti and Joe Incandela, Nobel prize for Higgs and Englert, 4 July 2012 Image source: CERN; using WLCG resources
GW150914, Nobel prize Rainer Weiss, Barry Barish, Kip Thorne, souces LIGO, Caltech, and MIT https://www.ligo.org/news.php; using OSG, select EGI sites, and REFEDS federated ID



http://manyworldstheory.com/2013/10/03/the-9-kinds-of-physics-seminar/

Q&A time!

David Groep, davidg@nikhef.nl
https://www.nikhef.nl/~davidg/presentations/
https://orcid.org/0000-0003-1026-6606
CC BY

Nikhef
Maastricht University | Department of Advanced Computing Sciences

DACS UM

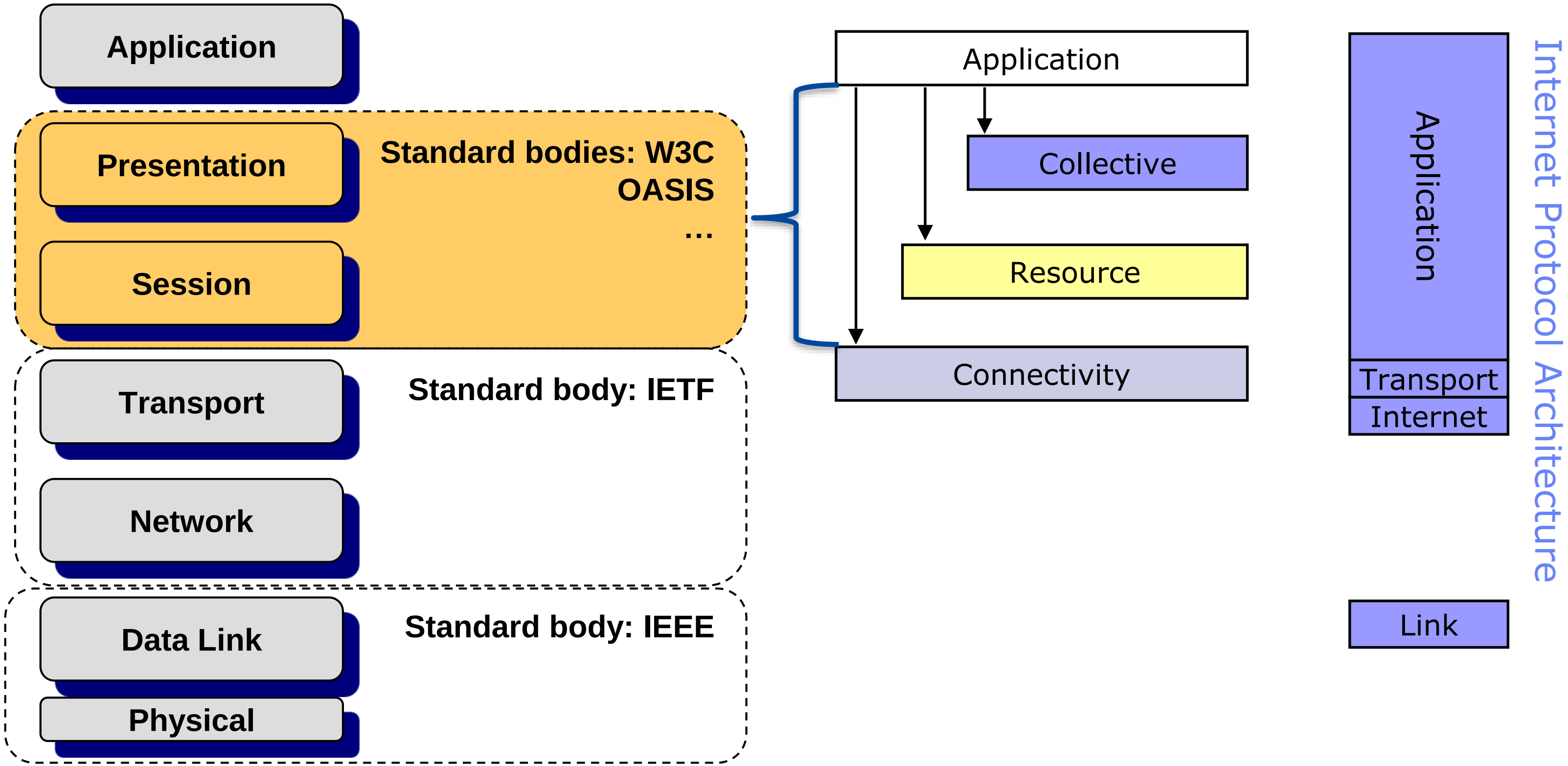
Ancillary materials

Open Systems Interconnection model (OSI model)

Layer			Function
Host layers	7	Application	High-level protocols (resource sharing, remote file access)
	6	Presentation	Translation of data between a networking service and an application
	5	Session	Managing communication sessions, i.e., continuous exchange of information in the form of multiple back-and-forth transmissions between two nodes
	4	Transport	Reliable transmission of data segments between points on a network
Media layers	3	Network	Addressing, routing and traffic control
	2	Data link	Transmission of data frames between two nodes connected by a physical layer
	1	Physical	Transmission and reception of raw bit streams over a physical medium

OSI X.200 layering model, ITU-T (CCITT), https://www.itu.int/rec/T-REC-X.200; image adapted from https://en.wikipedia.org/wiki/OSI_model

OSI vs Internet Protocol Architecture model



Private (direct) peerings to distribute traffic load

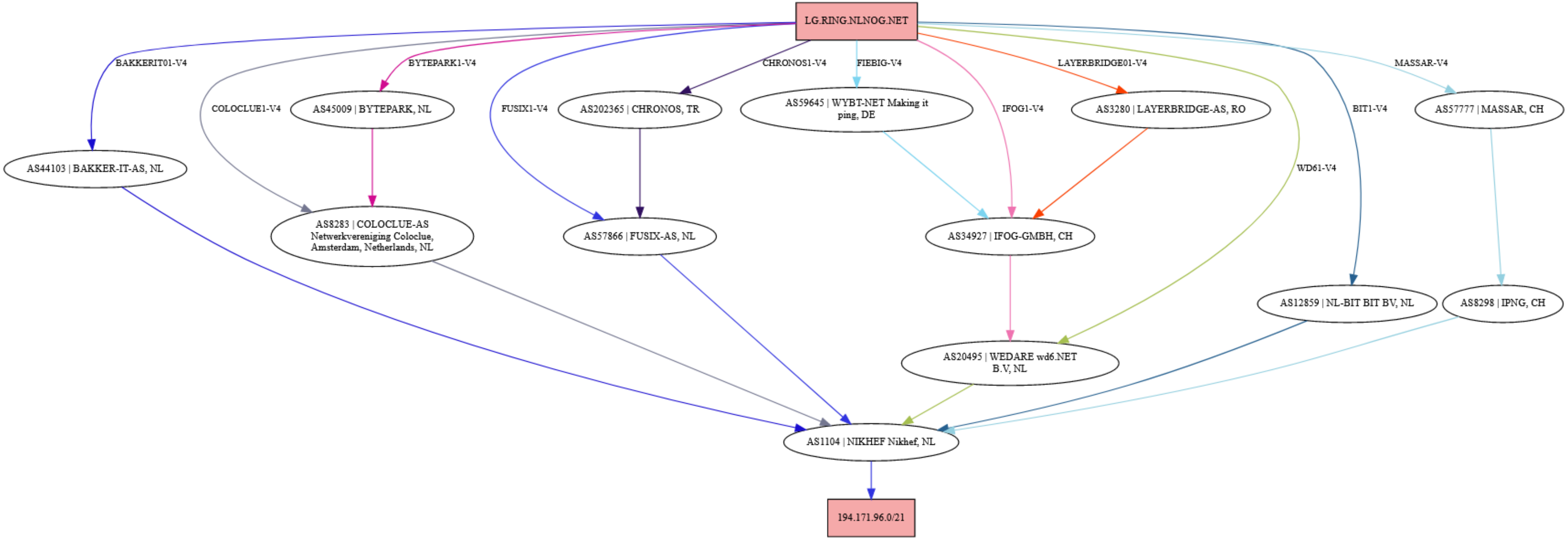
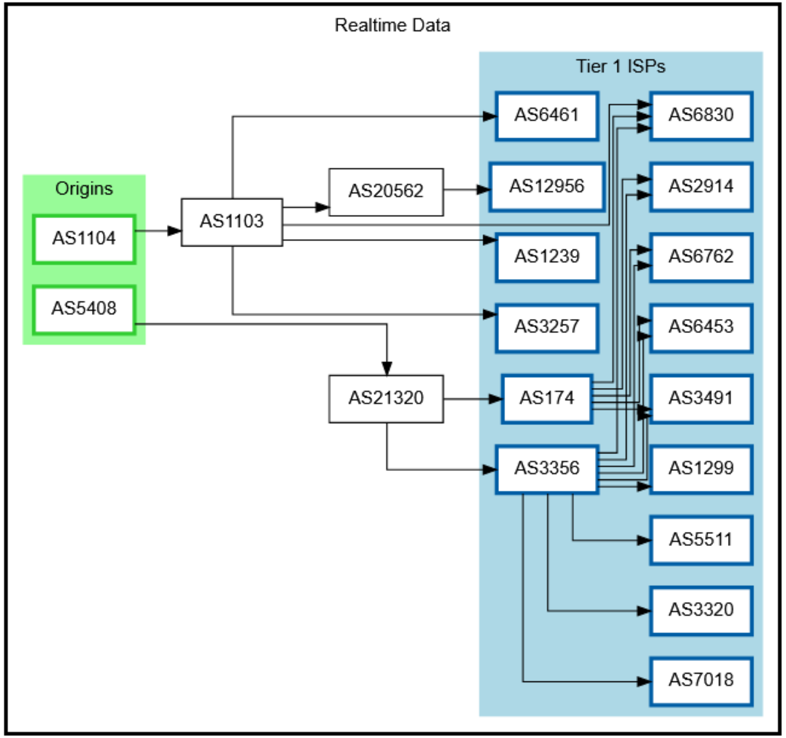


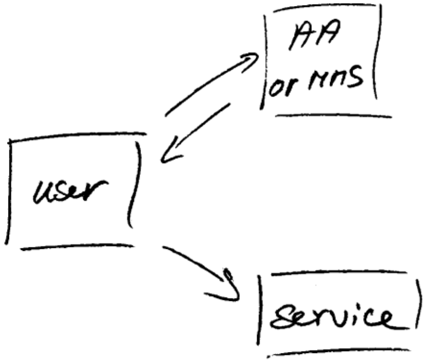
Image sources: NLNOG RING map https://lg.ring.nlnog.net/

Anycast – high availability leveraging BGP

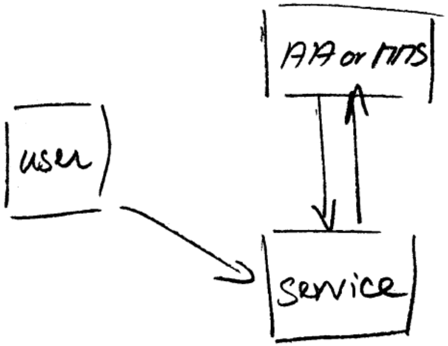


BGP.tools - https://bgp.tools/prefix/145.116.216.0/24#connectivity for anycasted RCauth.eu

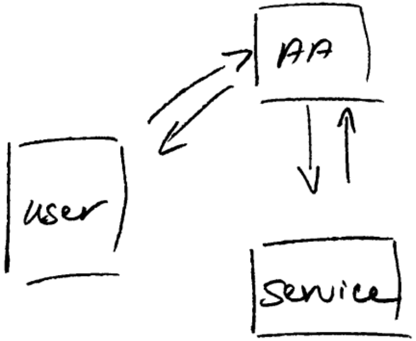
RFC2904 authorization models: three AuthZ flows



‘push’



‘pull’



‘agent’

Authorization models: AAA Authorization Framework, RFC2904, Vollbrecht et al.

OAuth2 & JWTs: assertions can be quite detailed

```
$ echo $AT | jwt
...
✻ Payload
{
  "wlcg.ver": "1.0",
  "sub": "a1b98335-9649-4fb0-961d-5a49ce108d49",
  "aud": "https://wlcg.cern.ch/jwt/v1/any",
  "nbf": 1593004542,
  "scope": "storage.read:/ storage.modify:/",
  "iss": "https://wlcg.cloud.cnaf.infn.it/",
  "exp": 1593008142,
  "iat": 1593004542,
  "jti": "da0a2f89-3cbf-42a7-9403-0b43d814551d",
  "client_id": "edfacfb1-f59d-44d0-9eb6-a745ac52f462"
}
```

OAuth2 Access Token following the WLCG AuthZ WG Profile, from: https://wlcg-authz-wg.github.io/wlcg-authz-docs/token-based-authorization/

Development background

Scope and structure